#1

C.5 0 6 8 5
4

PROGRESS REPORT SUMMARY	GRANT NUMBER LM06726-02	
PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR Norberto Ezquerra, Ph.D.	PERIOD COVERED BY THIS REPORT FROM 02/01/98	THROUGH 01/31/99
APPLICANT ORGANIZATION Georgia Institute of Technology		

TITLE OF PROJECT *(Repeat title shown in Item 1 on first page)*
Knowledge Discovery in Distributed Cardiac Imagebases

a. Human Subjects *(Complete Item 7 on the Face Page)*
Use of Human Subjects ☐ Change ☒ No Change Since Previous Submission

b. Vertebrate Animals *(Complete Item 8 on the Face Page)*
Use of Vertebrate Animals ☐ Change ☒ No Change Since Previous Submission

(SEE INSTRUCTIONS)

Questions to be answered: (1) Has there been a change in the other support of key personnel since the last reporting period? No. (2) Will there be, in the next budget period, significant rebudgeting of funds and/or change in the level of effort for key personnel from what was approved for this projects? No.

A. Specific Aims The specific aims have not changed: (i) to create mining algorithms to discover potentially new knowledge from cardiovascular perfusion images and other (non-image) patient data; (ii) to use the knowledge derived from the mining efforts to enrich a knowledge-based system, PERFEX, designed to interpret the medical imagery and other data; and (iii) to extend knowledge-based (KB) processing and mining methods to distributed environments.

B. Studies and Results Our first year of research resulted in 3 primary achievements: (1) Development of four computational elements necessary to support the major phases of the research: (i) design of a data format suitable for efficient mining of images and other clinical data; (ii) design and construction of a medical database (DB) consisting of these image and non-image data, (iii) creation of a DB management system (DBMS); and (iv) design and implementation of an architecture and middleware to support both distributed knowledge discovery and knowledge-based processing; (2) Design and implementation of innovative mining algorithms; and (3) Validation of the results generated by the mining algorithms with the clinical DB. These are briefly summarized below.

An architecture was designed and implemented that allows a remote (physician) "client" to submit his/her medical data for clinical interpretation using a JAVA-enabled WWW browser. The data thus submitted is translated into a prespecified format, and subsequently processed by the KB system PERFEX (within a few seconds). A report is produced presenting a medical interpretation of the patient data that can be queried by the user. Significantly, the same architecture serves to provide additional information for mining, as the files submitted by external clients augment the underlying, global DB. In addition, a search was conducted of the Emory U. Cardiac DB to identify patients who had undergone both perfusion (SPECT) studies and coronary angiography (the latter is used as a gold standard in the detection of coronary artery disease), yielding an initial set of 661 patient files containing both image and clinical patient information. The records contain nearly 150 fields that are quite comprehensive in terms of diagnostic significance. These clinical files have been converted into a new binary format that optimizes mining efficiency (files from external medical centers will be converted to this efficient format also). The DB combined with the new format (and translation programs) form the initial core of clinical information to be mined and to subsequently enrich the KB system with knowledge that may be new or confirmed with improved statistics.

An innovative mining algorithm was implemented that allows: an unlimited number of records (patient files), a high number of fields per record, flexible data input formats, and the addition or deletion of records. This permits incremental mining whereby the entire data set does not have to be mined again if a few records are added or deleted. Based on comparisons and re-use of previous mining results, incremental mining can save as much as 78% of

GENDER AND MINORITY INCLUSION
Provide the number of subjects enrolled in the study *to date* (cumulatively since the most recent competitive award) according to the following categories. (See Page 8 for definitions.) If there is more than one study, provide a separate table for each study. In addition, report on the subpopulations which are included in the study.

Study Title

	American Indian or Alaskan Native	Asian or Pacific Islander	Black, not of Hispanic Origin	Hispanic	White, not of Hispanic Origin	Other or Unknown	TOTAL
Female							
Male							
Unknown							
TOTAL							

computational time as suggested by the medical data thus far studied. The efficiency of the mining algorithm has been studied, showing that is of the order of the square of the number of associations (i.e., rules found), and efforts are already underway to further improve efficiency. In addition, a DBMS was developed to address the problem of the dynamic nature of the DB, since the data records can easily increase or decrease over time. A relational DB has been designed consisting of patient-related information, data-format (and translation) information, and information related to rules or associations resulting from mining operations; there are seven relations in this scheme (as will be reported in a publication under preparation). A 288 patient subset has been used to test and validate the data mining algorithm, which in turn served to also validate the new data format. The results of mining this dataset, as described earlier, showed that the performance (in particular the specificity) of the knowledge base can be improved through the inclusion of patient symptoms, EKG results, chest attenuation, and several other clinical information records. We are currently verifying new patterns that appear to be new knowledge, and expect to validate these findings very soon with the entire (larger) DB.

C. Significance The creation of a preliminary Web-based system that allows remote KB processing is a potentially powerful contribution since such a system would (a) permit a physician (or, more generally, a remote user) to quickly and easily obtain an interpretation of a patient case consisting of both image and other clinical information, and (b) permit this same patient record to be added to a global DB that will subsequently be mined to uncover potentially new associations or rules. Also, the creation of an incremental DB mining algorithm for finding associations is an important contribution not only because of the potential of uncovering new relationships in the medical data, but also because it allows the incremental addition or deletion of data records without having to redo much of the mining operations anew. Performance studies have also resulted in improvements in algorithm efficiency.

It should be noted that the thrust of our efforts during this first year of research has been placed on designing and constructing the underlying DB, DBMS, algorithms, and communications architecture, methods and tools, rather than on actual discovery of new knowledge or distributed KB processing. Although this may not appear to be a "spectacular" research result, it was actually an expected one, since the computational investigations performed in this first phase are foundational to, and required by, the subsequent KB processing and mining operations in distributed settings. In addition, it should be observed that the development and implementation of these informational and middleware methods and tools represent a significant effort in terms of iterative design, software engineering, algorithmic efficiency analysis, and sheer programming workload.

D. Plans for Next Funding Period The plans do not deviate from our proposed program, focusing on six research goals: (1) implementation of the mining algorithms with a large, local DB of 1000 patient cases (most of which has recently been created); (2) execution of incremental mining algorithms with the data records obtained via the Internet and derived from the five external sites collaborating on the project (as set forth in the proposal); (3) development of algorithms for mining time-sequence data (i.e., records which have changed over time); (4) analysis and validation of knowledge resulting from the mining (in terms of either new knowledge, inferences that confirm existing knowledge, or information that may be in conflict with currently existing knowledge); (5) creation of an interface that will facilitate the visualization and/or interpretation of mining results; (6) further refinement of the web-based version of PERFEX such that it becomes externally accessible as well as increasingly efficient and interactive; and (7) implementation of visualization and query operations in PERFEX to further enhance and support remote consultations. As results emerge from mining and remote knowledge-based processing, we plan to publish and disseminate the findings throughout the clinical, medical informatics, and database communities.

E. Publications

1."Knowledge-Guided Visualization of 3D Medical Imagery" to be published in Future Generations Computer Systems, Elsevier Science B.V., Special Issue on Scientific Visualization.
2. "Mining of Cardiovascular Medical Images" to appear in ACM SIGBIO Newsletter, special issue on Biomedical Knowledge Discovery and Data Mining.

As noted above, manuscripts are under preparation as results from mining and remote KB processing operations emerge.

F. Project-Generated Resources As the first year of research comes to an end, the main resources generated thus far are a medical DB especially formatted for mining, incremental mining algorithms, and middleware methods and tools supporting web-based consultations with the knowledge-based system PERFEX. Although these algorithms and tools are still undergoing iterative re-design and refinement (emphasizing efficiency and ease-of-use), we strongly feel that these efforts can potentially serve as a model for other distributed image mining and KB processing applications.

G. Inventions and Patents None.

Questions to be included with Form Page 5 (Progress Summary Report)

1. Has there been a change in the other support of key personnel since the last reporting period? No.

2. Will there be, in the next budget period, significant rebudgeting of funds from what was approved for this project? No.

3. Will there be, in the next budget period, a change in the level of effort for key personnel from what was approved for this project? No.

4. Is it anticipated that an estimated unobligated balance (including prior year carryover) will be greater than 25 percent of the current year's total budget? No.

CHECKLIST	GRANT NUMBER LM06726-02

1. ASSURANCES/CERTIFICATIONS *(See Instructions, Page 9)*

The following assurances/certifications are made and verified by the signature of the OFFICIAL SIGNING FOR APPLICANT ORGANIZATION on the FACE PAGE of the application. If unable to certify compliance where applicable, provide an explanation and place it after this page.

• Human Subjects; • Vertebrate Animals; • Debarment and Suspension; • Lobbying; • Delinquent Federal Debt; • Research Misconduct; • Civil Rights (Form HHS 441 or HHS 690); • Handicapped Individuals (Form HHS 641 or HHS 690); • Sex Discrimination (Form HHS 639-A or HHS 690); • Age Discrimination (Form HHS 680 or 690); • Financial Conflict of Interest.

2. PROGRAM INCOME *(See Instructions, Page 10)*

All applications must indicate whether program income is anticipated during the period(s) for which grant support is requested. If program income is anticipated, use the format below to reflect the amount and source(s).

Budget Period	Anticipated Amount	Source(s)
NONE	NONE	NONE

3. INDIRECT COSTS

Indicate the applicant organization's most recent indirect cost rate established with the appropriate DHHS Regional Office, or, in the case of forprofit organizations, the rate established with the appropriate PHS Agency Cost Advisory Office. Indirect costs will *not* be paid on foreign grants, construction grants, grants to Federal organizations, grants to individuals, and conference grants. Follow any additional instructions provided for Research Career Awards, Institutional National Research Service Awards, and specialized grant applications.

☐ DHHS Agreement dated: _____ ☐ No Indirect Costs Requested.

☐ No DHHS Agreement, but rate established with _____ Date _____

CALCULATION*

Entire proposed budget period:

Amount of base $ _____ x Rate applied _____ % = Indirect costs $ _____

Add to total direct costs from form page 2 and enter new total on FACE PAGE, Item 9b.

*Check appropriate box(es):

☐ Salary and wages base ☐ Modified total direct costs base ☐ Other base *(Explain below)*

☐ Off-site, other special rate, or more than one rate involved *(Explain below)*

Explanation *(Attach separate sheet, if necessary.)*:

#2

PROGRESS REPORT SUMMARY	GRANT NUMBER	
	LM06726-03	
PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR	PERIOD COVERED BY THIS REPORT	
Norberto Ezquerra, Ph.D.	FROM	THROUGH
APPLICANT ORGANIZATION		
Georgia Institute of Technology	02/01/0	01/31/01

TITLE OF PROJECT *(Repeat title shown in Item 1 on first page)*
Knowledge Discovery in Distributed Cardiac Imagebases

a. Human Subjects *(Complete Item 7 on the Face Page)*
Use of Human Subjects ☐ Change ☒ No Change Since Previous Submission

b. Vertebrate Animals *(Complete Item 8 on the Face Page)*
Use of Vertebrate Animals ☐ Change ☒ No Change Since Previous Submission

(SEE INSTRUCTIONS)

Has there been a change in the other support of key personnel? No.
Will there be, in the next budget period, significant rebudgeting of funds? No.
Will there be a change in the level of effort for key personnel? No.
Is an unobligated balance expected? No.

Research Accomplishments

(A) SPECIFIC AIMS

The overall objective of the research program remains to discover potentially new knowledge regarding the assessment of coronary artery disease (CAD) by mining image and non-image databases both locally and at remote sites. Similarly, the three specific aims have also remained the same, as discussed next.

(B AND C) STUDIES, RESULTS, AND SIGNIFICANCE

During the current budget year, there has been substantial progress toward these three aims. These achievements and their significance are as follows:

Aim #1: Knowledge Discovery. There were four main thrusts related to this specific aim:

(i) After pre-filtering our patient data base (as explained in (ii) below) to retain only parameters associated with stress perfusion image characteristics and the results of coronary angiography, we were able to generate close to 900 meaningful interpretation rules associating the location of stress perfusion defects to the jeopardized coronary vessel territory. As reported in [Coo99], this is perhaps the most medically meaningful accomplishment: the discovery of potentially new knowledge regarding the relationships that may exist between image data, non-image data, and the assessment of coronary artery disease.

(ii) Rule-Finding Algorithm: Our initial association-mining algorithm was originally developed to work on records consisting of binary attributes. However, this past year we obtained a more complete dataset wherein each record consists of 62 attributes of alphanumeric/numeric data that are clinically meaningful (e.g., patient symptoms, electrocardiographic information, etc; see the attached publications). For the mining algorithms to fully exploit the richness of the data, the data had to be translated to records with binary attributes. Thus, data-filtering techniques were developed to perform this translation and also to improve algorithm efficiency. The significance of this effort is the creation of an innovative procedure with which to both filter and translate multiply-valued data into meaningful subsets in a binary representation, as reported in [Ord99a]. Importantly, a user interface (UI) was designed to facilitate mining operations (Figure 1). This UI will allow the mining algorithms to be used directly by clinical users.

(iii) A Clustering Algorithm to Discover Interesting Subspaces of Multidimensional Data: In this work, we examined the problem of more efficiently handling records containing multiple numeric attributes. The basic concept is that potentially interesting patterns in such data can be discerned by finding clusters. We have developed a

GENDER AND MINORITY INCLUSION
Provide the number of subjects enrolled in the study ***to date*** (cumulatively since the most recent competitive award) according to the following categories. (See Page 9 for definitions.) If there is more than one study, provide a separate table for each study. In addition, report on the subpopulations which are included in the study.

Study Title

	American Indian or Alaskan Native	Asian or Pacific Islander	Black, not of Hispanic Origin	Hispanic	White, not of Hispanic Origin	Other or Unknown	TOTAL
Female							
Male							
Unknown							
TOTAL							

clustering algorithm to discover low- and high-density regions in subspaces of multidimensional data for mining. Regions are considered interesting when they have a minimum "volume" and involve some maximum number of dimensions. This innovative and fast algorithm discovers high density regions (clusters) and low density regions (outliers, negative clusters, holes, empty regions) at the same time, as reported in [Ord99b].

(iv) An Alternative Metric for Support in Mining Associations from Databases: In finding associations, support is used as an indicator as to whether an association is interesting. In this work, we introduced an alternative measure of interestingness called bond. We proved that the antimonotonicity property related to support also applies to bond. This is important, because it allows the mining program to prune the search space when looking for interesting associations. We also proved that if associations have a minimum bond, then those associations will have a given lower bound on their minimum support and the rules produced from those associations will have a given lower bound on their minimum confidence as well. This basic contribution is described in [Omi99].

Aim #2: Knowledge Base (KB) Enrichment. The results of the mining operations are cast in the form of rules that are subsequently validated and introduced into our KB system, PERFEX, which is designed to interpret stress/rest myocardial perfusion. This system was validated using 655 prospective patient studies from Emory University [Gar99]. The certainty factors used by PERFEX, which reflect the severity of the perfusion defect, were modified to be an adjustable, more accurate variable. This analysis suggested that, in some cases, a change in the value of this factor resulted in no statistically significant differences between the automatic interpretations made by PERFEX and those made by human experts when using coronary angiography as a gold standard. Importantly, in other cases PERFEX actually showed superior performance relative to the human experts. The significance of this work is that the results of mining can improve the statistical (and probabilistic) foundations of rule-based processing, and that mining can also uncover knowledge that can be validated and incorporated in a KB system to consistently accurately interpret the medical information [Gar99].

Aim #3: Distributed Knowledge Discovery. We have begun to collect 400 patient studies from four collaborating sites (100 cases from each site) that will serve as our remote facilities for validation and evaluation, as proposed. These sites are Dr. Jack Ziffer's lab at Miami Baptist in Florida, Dr. Gordon DePuey at Roosevel-St-Lukes in NYC, Dr. Timothy Bateman at Cardiovascular Consultants in Kansas City and Dr. Jaume Candell Riera at Hospital Vall D'Hebron in Barcelona, Spain. This effort consists of (a) transmitting both the image data sets and textual data for each patient, and (b) translating both the textual DB records as well as the imagery (the latter also requiring pre-processing) to a predetermined format. We will have all 400 patient studies on site by January 2000 and totally translated and ready to use subsequently. In addition to these efforts, a test demonstrating remote access to PERFEX (a Web-based version of PERFEX) has been performed.

(D) PLANS FOR NEXT YEAR OF SUPPORT

During the next project period, we will continue to collect patient data both locally and at remote sites. Aims (2) and (3) will see increased activity during this period, as originally proposed. The translation and filtering methods will be further improved to handle heterogeneity of the DBs from different sites. It is likely that the increased number of records will not only help to discover potentially new knowledge, but may also uncover site-dependent biases and other trends. Algorithmic improvements such as clustering, bond, and filtering will be further explored. Importantly, with the introduction of the user interface, the mining algorithm will be placed directly in the hands of clinical users. As the data filtering and mining algorithms become more robust, and usable, and as more data records become available, it is expected that mining results will increasingly lead to the discovery of interesting knowledge. Thus, a major thrust of the next period will be to prepare and submit numerous manuscripts on medical knowledge discovery and validation, efficient mining algorithms, and Web-based image interpretation.

(E) PUBLICATIONS

Coo99 Cooke CD, Ordonez C, Garcia EV, Omiecinski E, Krawczynska EG, Folks RD, Santana CA, de Braal L, Ezquerra N: Data mining of large myocardial perfusion SPECT databases to improve diagnostic decision making. J Nucl Med 1999, 40 (5), p292P

Gar99 Garcia EV, Cooke CD, Folks RD, Santana CA, Krawczynska EG, Ezquerra NF, Vansant JP, Ziffer JA,: Expert system interpretation of myocardial perfusion tomograms: validation using 655 prospective patients. J Nucl Med 1999, 40 (5), p126P

Omi99 E. Omiecinski, "An Alternative for Support in Mining Associations in Databases" (manuscript to be submitted for publication) (1999).

Ord99a Carlos Ordonez and Edward Omiecinski, "Discovering Association Rules based on Image Content," IEEE Advances in Digital Libraries Conference, May (1999).

Ord99b C. Ordonez, E. Omiecinski, S. Navathe and N. Ezquerra, "A Clustering Algorithm to Discover Low and High Density Hyper-Rectangles in Subspaces of Multidimensional Data," Georgia Tech Technical Report: GIT-CC-99-20 (to be submitted for publication). (1999)

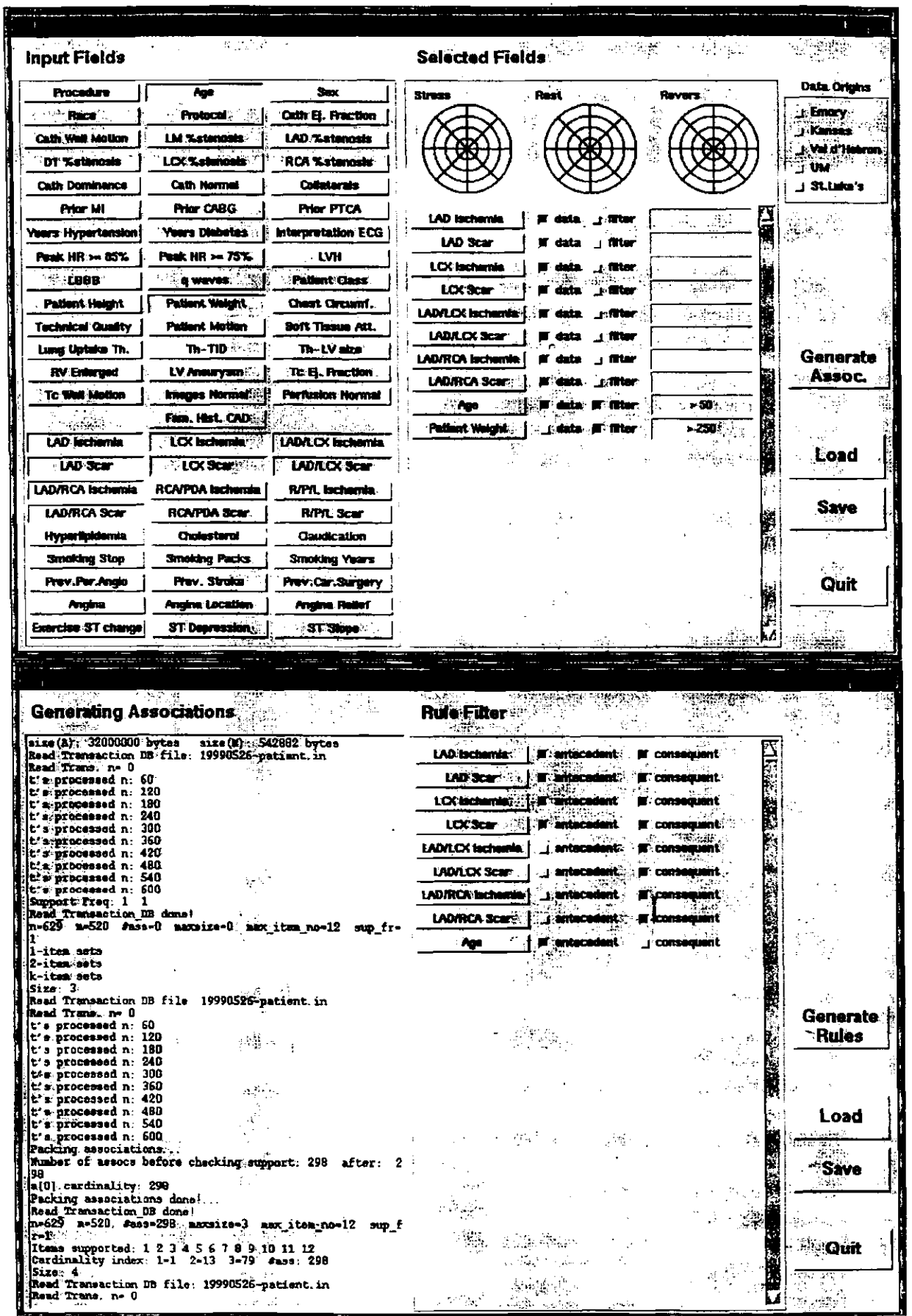

Figure 1: GUI example screens: interactive viewing, selection, and filtering of patient data.

CHECKLIST	GRANT NUMBER LM06726-03

1. ASSURANCES/CERTIFICATIONS *(See Instructions, Page 10)*

The following assurances/certifications are made and verified by the signature of the OFFICIAL SIGNING FOR APPLICANT ORGANIZATION on the FACE PAGE of the application. If unable to certify compliance where applicable, provide an explanation and place it after this page.

• Human Subjects; • Vertebrate Animals; • Debarment and Suspension; • Lobbying; • Delinquent Federal Debt; • Research Misconduct; • Civil Rights (Form HHS 441 or HHS 690); • Handicapped Individuals (Form HHS 641 or HHS 690); • Sex Discrimination (Form HHS 639-A or HHS 690); • Age Discrimination (Form HHS 680 or 690); • Financial Conflict of Interest.

2. PROGRAM INCOME *(See Instructions, Page 10)*

All applications must indicate whether program income is anticipated during the period(s) for which grant support is requested. If program income is anticipated, use the format below to reflect the amount and source(s).

Budget Period	Anticipated Amount	Source(s)

3. FACILITIES AND ADMINISTRATION (F & A) COSTS

Indicate the applicant organization's most recent F&A cost rate established with the appropriate DHHS Regional Office, or, in the case of forprofit organizations, the rate established with the appropriate PHS Agency Cost Advisory Office. F&A costs will ***not*** be paid on foreign grants, construction grants, grants to Federal organizations, grants to individuals, and conference grants. Follow any additional instructions provided for Research Career Awards, Institutional National Research Service Awards, and specialized grant applications.

☐ DHHS Agreement dated: _____ ☐ No F&A Costs Requested.

☐ No DHHS Agreement, but rate established with _____ Date _____

CALCULATION*

Entire proposed budget period:

Amount of base $ _____ x Rate applied _____ % = F&A costs $ _____

Add to total direct costs from form page 2 and enter new total on FACE PAGE, Item 9b.

*Check appropriate box(es):

☐ Salary and wages base ☐ Modified total direct costs base ☐ Other base *(Explain below)*

☐ Off-site, other special rate, or more than one rate involved *(Explain below)*

Explanation *(Attach separate sheet, if necessary.)*:

EMORY UNIVERSITY
Office of Sponsored Programs

1784 N. Decatur Road, Suite 510
Atlanta, Georgia 30322
Phone: 404/727-2503
Fax: 404/727-2397 or 2509
E-mail: osp@emory.edu
http://www.osp.emory.edu

November 11, 1999

Kathy Bean
Georgia Institute of Technology
801 Atlantic Drive
College of Computing
Atlanta, GA 30332-0280

RE: NIH Application
Principal Investigator (Emory) – Dr. Ernest Garcia

Dear Ms. Bean:

The purpose of this letter is to inform you that the appropriate programmatic and administrative personnel at Emory University involved in this grant application are aware of the PHS consortium policy and are prepared to establish the necessary inter-institutional agreement consistent with that policy.

If you have any questions or concerns please feel free to contact me at (404) 727-2503.

Sincerely,

Nancy L. Wilkinson, MPH
Assistant V.P. for Research

DETAILED BUDGET FOR INITIAL BUDGET PERIOD
DIRECT COSTS ONLY

FROM 2/1/00 THROUGH 01/31/01

PERSONNEL *(Applicant organization only)*

DOLLAR AMOUNT REQUESTED *(omit cents)*

NAME	ROLE ON PROJECT	TYPE APPT. *(months)*	% EFFOR ON PROJ	INST BASE SALAR	SALARY REQUESTED	FRINGE BENEFITS	TOTALS
Ernest Garcia	Principal Investigator	12	12%	125,900	15,221.	3,257.	18,479.
E. Krawczynska	Co-Inv	12	15%	41,140	6,217.	1,331.	7,548.
C. David Cooke	Computer Scientist	12	30%	56,187	16,983.	3,634.	20,617.
Russell Folks	Research Technologist	12	29%	52,885	15,452.	3,307.	18,758.
			0%		0.	0.	0.
			0%		0.	0.	0.
			0%		0.	0.	0.
			0%		0.	0.	0.
	SUBTOTALS →				53,873.	11,529.	65,402.
CONSULTANT COSTS						0.	
						0.	
						0.	0.
EQUIPMENT *(Itemize)*		0.				0.	
		0.				0.	
		0.				0.	
		0.				0.	
		0.				0.	
SUPPLIES *(Itemize by category)*							
Computer Supplies						250.	
Photographic Supplies						250.	
		0.				0.	
		0.				0.	
		0.				0.	
		0.				0.	500.
TRAVEL						0.	
1 Trip/year						1,158.	1,158.
PATIENT CARE COSTS	INPATIENT					0.	
	OUTPATIENT					0.	0.
ALTERATIONS AND RENOVATIONS *(Itemize by category)*							
						0.	0.
OTHER EXPENSES *(Itemize by category)*						0.	
Software maintenance for Blaze Expert		(60% allocation of $3,749)				$2,249	
		0.				0.	
		0.				0.	
		0.				0.	2,249
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD						**$**	**69,309.**
CONSORTIUM/CONTRACTUAL COSTS	DIRECT COSTS					0	0.
	FACILITIES AND ADMINISTRATION COSTS					$37,080	37,080.
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD *(Item 7a, Face Page)* →						**$**	**106,389.**

BUDGET JUSTIFICATION

Personnel:

Ernest V. Garcia Ph.D. - is the Principal investigator of the Emory subcontract. Dr. Garcia is a physicist with 22 years experience in the development of imaging algorithms particularly in nuclear cardiology. Dr. Garcia has been responsible for the development of computer methods for quantifying planar and tomographic myocardial perfusion studies that are currently in use in over 3,000 institutions world-wide. Dr. Garcia has had a major role in the development, justification and validation of the heuristic rules in the PERFEX expert system. Dr. Garcia's role in this proposal is to give clinical and scientific guidance, to assist in the design of algorithms, neural nets, expert systems and all aspects of the validation. He will be responsible for the Emory budget and personnel. He is scheduled to be funded for 12% of his time.

Elzbieta Krawczynska M.D., Ph.D. - is a co-investigator and an Assistant Professor in Radiology. She is a cardiologist with extensive clinical experience as well as experience in nuclear cardiology procedures. Her role will be to provide assurance that all methodology is consistent with the clinical aspects of nuclear cardiology. She will perform all data-base analysis, tabulation of results and will also serve as an expert observer in interpreting patient studies. She is scheduled to be funded for 15% of her time.

C. David Cooke, MSEE - is a computer scientist and an Assistant Professor of Radiology. He is an electrical engineer with extensive experience in computer hardware and software. In particular he has developed an expertise in the development environment of expert system shells. His role will be to develop software to communicate between medical imaging files and all forms of systems input. He will also perform all additional knowledge base and database mining algorithm implementation with clinical data. He will be the main developer of all algorithms generated by the personnel from this subcontract. He is scheduled to be funded for 30% of his time.

Russell Folks, R.T. - is a nuclear medicine research technologist with 9 years experience in research methods. His role will be in the processing patient studies (for research purposes), analysis, photographing and storing all of the studies used to develop and validate the methods described in this proposal. He is scheduled to be funded for 29% of his time.

Supplies

Computer supplies:

$250/yr is requested for the following computer supplies dedicated to this project: 1) Five 8mm tapes ($55), 2) 5 Zip disks ($50) are also requested to store our processed research studies 3) $145 for 150 sheets of color paper to record our research imaging results on printed paper for hard copy storage and for publications (we use about 400 sheets per year), The supplies requested are used in the Medical Development Imaging Center specifically for the research projects in this proposal. Since this project is almost exclusively for research and development of imaging database techniques and images take up a large amount of storage, there is a moderate amount of storage requirement.

Photographic Supplies:

$250/yr is requested for film to record original data, for slides and video tapes associated with this research.

Travel:

$1,158 is requested to pay in part for one trip to present results from these investigations at clinical/technical meetings or to interact with one of the collaborating sites.

Other expenses -

Software maintenance: A 60% allocation of the software maintenance for the Blaze expert software is requested. This is the software that runs PERFEX. The expert system being developed and implemented on the WEB in this project.

OTHER SUPPORT

Garcia, Ernest V

ACTIVE

HL42052-11 (Garcia)	12/01/88-12/31/99	Garcia 20% effort
NIH/NHLBI	$233,896 (direct)	

A unified approach to quantify and visualize cardiac imagery

The major goal of this project is to multi-dimensionally quantify, unify, and visualize the coronary arterial tree and the myocardial perfusion distribution. No overlap with this grant.

Cooke 25% effort, Folks 23% effort, 0% effort for others

NO GRANTS PENDING

Renewal of HL40252 is pending.

Volume 40, Number 5 (Supplement) • May 1999

SNM 46th Annual Meeting
Los Angeles, California
June 6-10, 1999

JNM

Official Publication of
the Society of Nuclear Medicine

No. 1290

DATA MINING OF LARGE MYOCARDIAL PERFUSION SPECT (MPS) DATABASES TO IMPROVE DIAGNOSTIC DECISION MAKING. C. D. Cooke*, C. Ordonez, E. V. Garcia, E. Omiecinski, E. G. Krawczynska, R. D. Folks, C. A. Santana, L. DeBraal, N. F. Ezquerra, Emory University, Atlanta, GA; Georgia Institute of Technology, Atlanta, GA. (101208)

Data mining is the automated discovery of unknown, nontrivial and potentially useful information from large databases. **Objective:** To discover associations between textual/imaging data in large MPS databases to be used for improving diagnostic decision making by either human experts or expert systems. **Methods:** A database was generated from 654 patients who had undergone both stress/rest MPS and coronary angiography (CA). The database combined textual patient and CA information from the cardiac data bank with imaging information from the output of the CEqual program for quantifying MPS. These files were converted to a format where each field became a binary variable as required by the data mining algorithm. For each patient, this resulted in 475 separate binary variables for the textual fields and 64 per perfusion defect. Two approaches were used to reduce the huge number of possible associations (IF & THEN rules) found by the system. First, a field filtering program was used to focus the associations to specific fields such as the relation "IF MPS defect THEN CA stenosis". Second the associations were limited to at least a 20% *support* (statistical significance) and a 90% *confidence* (rule strength). These variables were then input to a novel data mining algorithm which significantly reduced I/O and CPU overhead. **Results:** Initial mining has resulted in 950 rules and their associated confidences. We are in the process of comparing these rules to those in our previously developed expert system for interpreting MPS and in converting the confidence found to certainty factors used by the expert system for an expected overall improvement in diagnostic accuracy. **Conclusion:** Data mining of large databases of combined textual and imaging information promises to improve the diagnostic accuracy of human experts and expert systems.

No. 1291

DESCRIPTION AND PERFORMANCE OF A NUCLEAR MEDICINE RELATIONAL DATABASE MANAGEMENT SYSTEM FOR MINING PATIENT FOLLOW-UP. W. E. Gainey*, S. J. Cullom, K. L. Moutray, J. A. Case, J. H. Okeefe, A. I. McGhie, T. M. Bateman, Mid America Heart Institute, Kansas City, MO; Cardiovascular Consultants, P.C., Kansas City, MO. (101317)

Introduction: Nuclear Medicine databases are increasingly utilized for characterizing patient management patterns. Flat File Databases (FFDB),commonly used in nuclear medicine departments,only permit 1 to 1 association of a single study with a finite number of follow-up events, many which are redundant. A Relational Database Management System (RDBMS) is a generalized approach to data mining which permits a 1 to Many relationship with an arbitrary number of follow-up fields by utilizing unique data identifiers. Follow-up is defined as additional studies from external databases PTCA,CABG, CATH,and professional intervention via questionnaire. We have implemented an RDBMS into a large nuclear department and compared the performance of the existing FFDB for determination of patient follow-up. ***Methodology:*** Searchable data for this study included such follow-up patient information variables as subsequent studies and procedures, professional interaction, and major events (i.e. MI,Death). Each database was queried for follow-up on the same randomly selected 1000 patients. The representative sample of findings was analyzed for % of patients with data redundancy in follow-up, and total number of unique follow-up events. ***Conclusion:*** This study illustrates that a RDBMS eliminated redundancy in patient follow-up compared to the traditional methods of data mining.

Results (p<0.05)

	Redundancy in Pt. Follow-Up	
	Patient (%)	Events
FFDB	15.1	151
RDBMS	0	0

No. 1292

CAN A VOICE RECOGNITION DICTATION SYSTEM FOR NUCLEAR MEDICINE REPORTS BE SUCCESSFULLY USED IN AN ACADEMIC HOSPITAL SETTING? L. M. Fig*, B. Shapiro, R. S. Steventon, M. D. Gross, Veterans Administration Medical Center, Ann Arbor, MI. (500455)

Objectives: Prompt delivery of accurately transcribed Nuclear Medicine (NM) reports is the goal of every NM department. Computer systems that convert dictation directly into text are becoming more widely available. The purpose of this study was to determine the suitability of a commercial voice activated dictation system (VADS) for implementation in a VA teaching NM department where radiology/NM residents with highly variable clinical experience/training and speech accent rotate weekly. Methods: 1) The accuracy of the VADS (MedSpeak, IBM) was evaluated in a simulated clinical setting: 13 residents (8M, 5F) and 3 staff NM physicians without prior exposure to VADS dictated standardized reports which contained typical wording for cardiac, bone and lung V/Q studies. Initial dictations using the standard (non-enrolled) voice recognition program were graded for accuracy and expressed as % words incorrectly transcribed. 2) Personal voice enrollments were performed for inaccuracy >5% and the dictation repeated. 3) Turnaround time for clinical report completion was assessed for comparable periods before and after implementation of the VADS. Results: Overall, mean inaccuracy of 10 subjects with standard US accents was 5.9% and for 5 with non-US accents was 11.25%. Inaccuracy was significantly greater for women residents than men for both US and non-US accents (11.9% vs. 3.4% for US accents and 22.95% vs. 8.325% for non-US accents). Following personal enrollment mean overall inaccuracy fell from 7.75% to 3.9%. First day individual training took a mean of 47 min (without personal voice enrollment) plus 60 min for enrollment; additional support was required for 2-3 days to answer specific questions (total estimated time 30 min/resident/day). Report turnaround time was reduced by 51.2% after introduction of VADS. Conclusions: VADS can be successfully integrated into an academic multiuser NM setting. These systems may reduce report turnaround time but serious consideration must be given to the time/personnel effort involved in: 1) adequate training of rotating residents and 2) personal enrollment of women's voices and non-US accents.

No. 1293

SIMPLE LOW-COST VARIABLE-RESOLUTION TELERADIOLOGY SYSTEM. C. K. Hoh*, J. Darling, V. J. Neal, J. Czernin, Long Beach Community Medical Center, Long Beach, CA; University of California at Los Angeles Medical Center, Los Angeles, CA. (100824)

To provide coverage of emergency and routine nuclear medicine studies at affiliated hospitals with a distance of over 30 miles from UCLA, a simple teleradiology system was implemented. A large flat bed transparency scanner (12.2 x 17.2") connected to a personal computer (PC) was installed at each site. To scan in the images, the technologist uses software supplied with the scanner, Photoshop (Adobe Systems Inc.). Alternatively, for technologist unfamiliar with the system, the scanner can be entirely controlled by the interpreting physician using remote control software, pcANYWHERE (Symantec). The remote control also allows the physicians to rescan a portion of the image using a higher resolution (up to 1200 dpi). Images were saved as JPEG files and transfered using pcANYWHERE via a 28K modem to home PCs or mobile/cellular PCs. For a 11x14" CXR scanned in at 200 dpi, 2 minutes were required for a file transfer. Adobe Photoshop was used as the image display software at the home and mobile sites. Since 4/98, over 140 images were transmitted, which included V/Q scan (& CXRs)(79%), GI bleeding, HIDA, WBC, and bone scans. Image quality was excellent and there was no difference between the teleradiology image interpretations and the final reports based on the actual film readings. A simple teleradiology system is feasible using a large flat bed transparency scanner. Since the system is based on film, there is no conversion of proprietary image file formats; however, no cine or gated images can be viewed. This system has: 1) no compromise in diagnostic image quality, 2) capability for all imaging modalities, 3) total system cost of less than $5000 per site, 4) no special requirements for software programming or service support, 5) no special training for remote site technologists, 6) secure file transfers using a regular phone line.

stress ratio was the ress ratio
<45% optimized id speci-
ficity=96%). Appl up of 58
ICMP and 31 NCM... l a sensi-
tivity=42% and specificity=90%. **Conclusions:** Presence of a severe stress myocardial perfusion defect is a hallmark of ICMP. An objective value of a stress ratio <45% is a very specific marker for differentiating ICMP from NCMP by SPECT.

Patient Group (n)	LVEF (%)	EDV (ml)	Stress Study Ratio	Stress Study Extent	Rest Study Ratio	Rest Study Extent
NCMP (63)	25±9	188±74	51±4	219±190	51±5	222±160
ICMP (81)	26±8	182±66	41±7*	603±339*	42±7*	529±346*

EDV=end-diastolic volume; *p<1x10^{-3} vs NCM

No. 508

EXPERT SYSTEM (PERFEX) INTERPRETATION OF MYOCARDIAL PERFUSION TOMOGRAMS: VALIDATION USING 655 PROSPECTIVE PATIENTS.

E. V. Garcia*, C. D. Cooke, R. D. Folks, C. A. Santana, E. G. Krawczynska, N. F. Ezquerra, J. P. Vansant, J. A. Ziffer, Emory University, Atlanta, GA; Georgia Institute of Technology, Atlanta, GA. (100761)

Objective: To validate PERFEX in identifying the presence (P) and location (L) of CAD using 655 stress/rest myocardial perfusion prospective SPECT studies in patients who also underwent coronary arteriography (cath). **Methods:** Using data from 461 other patient studies we implemented and refined 253 heuristic rules that best correlated the P and L of perfusion defects (PDs) on SPECT studies with cath demonstrated CAD and with expert visual interpretations. The PDs were identified from polar CEqual maps as pixels with counts below gender-matched normal limits. PERFEX uses the certainty of the L, size, shape and reversibility of the PDs to infer the certainty of the P and L of CAD. The patient population was comprised of 480 CAD patients and 175 normals, 449 were males and 206 females. The visual interpretations (V) of slices and maps, vessel stenosis from cath (C) and PERFEX (Px) interpretations were all accessed automatically from data bases and used to automatically generate inter comparisons as shown in the table below. **Results:** PERFEX demonstrated a significantly higher sensitivity and lower specificity than visual interpretation for identifying the P and L of CAD (*p<.05 vs cath). **Conclusion:** PERFEX provides fast (<15secs) and clinically useful interpretations and justifications.

	CAD		LAD		LCX		RCA	
	SN*	SP	SN*	SP*	SN*	SP*	SN*	SP*
V vs C	87	21	69	59	61	86	73	71
Px vs C	93	18	84	27	76	47	82	45
Px vs V	94	34	91	38	94	60	90	60

No. 509

VARIABILITY OF MYOCARDIAL PERFUSION SPECT: CONTRIBUTION OF REPETITIVE PROCESSING, ACQUISTION, AND TESTING.

L. A. MacDonald*, M. D. Elliott, S. M. Leonard, M. A. Parker, M. W. Groch, S. M. Spies, R. C. Hendel, Northwestern University, Chicago, IL. (500523)

Background: Although tomographic perfusion imaging is a mainstay in the evaluation of pts with known or suspected coronary artery disease, the variability (V) of the technique is poorly defined, which is especially pertinent for serial imaging. Methods: Accordingly, 10 pts underwent repeated stress and rest SPECT; all pts underwent both stress and rest imaging on separate days (inter-test V; INTER). An additional acquisition was also obtained (same day) for each of the stress and rest studies (intra-test V; INTRA). Furthermore, each acquitision was reprocessed. Polar plots were then constructed for each data set and divided into 9 regions (septal, anterior, inferior and lateral regions in the base and mid-cavity, apical region). The mean activity per region calculated as percentage of maximum activity in the image. The processing V, INTRA, and INTER was then determined by comparing each corresponding image pair and represented as the standard error of the measurement (SEM). Results: The SEM is displayed in the table for INTRA and INTER. INTRA accounted for 99% and 73% of the stress and rest INTER, respectively. Repeat processing had a SEM of 1.87, which accounts for about one-half of the observed V. The stricted to the myocardium distal to the membranous septum, a marked duction in V was noted, with the INTER declining form 5.01 to 3.77 in resting studies. Conclusions: Image processing accounts for a substan amount of the V noted in serial imaging studies. An additional import factor in V is due to pt positioning and image acquisition (INTRA). Tr cating quantitation at the begining of septal dropout substantially redu V. Advances in automated processing may help to reduce V. Establish the limits of test reproducibility may help define true changes in se SPECT imaging.

	INTRA	INTER
STRESS	4.04	4.05
REST	3.65	5.01

No. 510

MYOCARDIAL SPECT MOTION ARTIFACT AS FUNCTION OF MOTION TYPE AND NUMBER (CAMERA DETECTORS..

D. A. Hillier*, J. W. Wall Mallinckrodt Institute of Radiology, St. Louis, MO. (500559)

Objectives: Prior myocardial SPECT studies on the appearance a severity of motion artifact have assumed use of a single head camera, lized patient data, and nearly always assume a single motion. This stu assesses pure motion artifact with phantom data of single non-return vs. multiple motions and the implications for camera head number. I dual-detector acquisition, the acquisition time is halved compared with single head case, but the motion pattern is different, with more moti transitions (one for each head and at the boundary of the data between two heads); it is thus unclear which would yield the greatest artifact. Me ods: In t study, 1-3 pixel vertical motion was simulated at various tin points using projections from a Data Spectrum phantom. A 180o orbit ing a single detector vs. a 90o dual detector system was evaluated. Artif severity was quantified from bullseye images. Results: Artifact sever varied with the degree of motion, being of moderate severity with a 2-pi (1.2-cm) motion, worst during mid-acquisition, and nearly identical in t two orbits. If a single motion is equally likely to occur during any time p riod during the study, then the probability of significant artifact is propo tional to the area under the quantitated artifact severity vs. time curv The peaks are similar, but the single-head curve is broader (since exa time is longer) and the probability, therefore, of clinically significant m tion artifact is higher. Apical defect artifacts were most prominent. Mul ple repeated motion consisting of alternating 2-pixel up and down retu ing motions (a common type of motion in our experience), also assesse produces relatively little artifact (volume averaging effects, but no sha discontinuities, are seen). Conclusion: The degree of single-motion ar fact is the same in single- and dual-head acquisitions. Due to longer a quisition time, more frequent motion artifact will likely be seen in the si gle-head case. Multiple repetitive motion produces less artifact than sing motion in either acquisition.

No. 511

THREE-DIMENSIONAL (3D) COLOR MODULATE DISPLAY OF MYOCARDIAL SPECT PERFUSIO DISTRIBUTIONS ACCURATELY ASSESSES CORONAR ARTERY DISEASE (CAD).

C. A. Santana*, E. V. Garcia, J. Vansant, E. G. Krawczynska, R. D. Folks, C. D. Cooke, J. Ziffer, Emory University, Atlanta, GA. (100766)

Objectives: To assess the usefulness of the visual assessment of 3D d plays (color modulated to count density)in comparison to oblique tom grams (slices) of myocardial SPECT perfusion distributions in the dete tion and localization of CAD. **Methods:** Sixty-two consecutive patien (61 ± 11 years,21 women) who had undergone conventional dual isoto perfusion SPECT were retrospectively chosen, 50 had previous corona arteriography (cath)(42 with at least 1 stenosis ≥50%, 8 "normals", with previous MIs) and 12 had < 5% likelihood of CAD. Three reade visually interpreted the 3D displays and slices in separate sessions blind to: a) their previous readings, b)interpretation of others, c)cath result Readers used a 5 point score for their interpretations. Their average sco was used for both ROC analysis and comparison of the accuracy of t techniques. Sensitivity (SN), specificity (SP), Normalcy Rate (NR) ar area under the ROC curves (Az) were determined for the detection CAD and localization in vascular territories (vessels). **Results:** The tab

An Alternative for Support in Mining Associations in Databases

Edward Omiecinski *

College of Computing
Georgia Institute of Technology
Atlanta, GA 30332

e-mail: edwardo@cc.gatech.edu

Abstract

Data mining is defined as the process of discovering significant and potentially useful patterns in large volumes of data. Discovering associations between items in a large database is one such data mining activity. In finding associations, support is used as an indicator as to whether an association is interesting. In this paper we introduce an alternative measure of interestingness called bond. We prove that the important antimonotonicity property applies to bond. We also prove that if associations have a minimimum bond, then those associations will have a given lower bound on their mimimum support and the rules produced from those associations will have a given lower bound on their minimum confidence as well. Although associations that have that mimimum support (and likewise their rules that have minimum confidence) may not satisfy the minimum bond constraint. We describe the algorithm that efficiently finds all associations with a minimum bond and present some experimental results.

1 Introduction

The past few years has seen a tremendous interest in the area of data mining. Data mining is generally thought of as the process of finding hidden, nontrivial and previously unknown information in a large collection of data [13]. Exploiting large volumes of data for superior decision making by looking for interesting patterns in the data has become a main task in today's business environment. In particular, finding associations between items in a database of customer transactions, such as the sales data collected at super market check out counters [1, 2, 6, 7, 9, 12, 15, 16, 17] has become an important data mining task. Association rules identify items that are most often bought along with certain other items by a significant fraction of the customers. For example, we may find that "95 % of the

*This work was supported in part by Grant LM 06726-02 from the National Library of Medicine.

customers who bought bread also bought milk." A rule may contain more than one item in the antecedent and the consequent of the rule. Every rule must satisfy two user specified constraints: one is a measure of statistical significance called *support* and the other a measure of goodness of the rule called *confidence.*

Formally, the problem can be stated as follows [1, 2]: Let $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$ be a set of m distinct literals called *items.* $\mathcal{D}$ is a set of variable length transactions over $\mathcal{I}$. Each transaction *contains* a set of items $i_i, i_j, \ldots, i_k \subset \mathcal{I}$. A transaction also has an associated unique identifier called TID. An *association rule* is an implication of the form $X \longrightarrow Y$, where $X, Y \subset \mathcal{I}$, and $X \cap Y = \emptyset$. X is called the antecedent and Y is called the consequent of the rule.

In general, a set of items (such as the antecedent or the consequent of a rule) is called an *itemset.* The number of items in an itemset is called the *length* of an itemset. Itemsets of some length k are referred to as k–itemsets. For an itemset $X \cdot Y$, if Y is an m–itemset then Y is called an *m-extension* of X.

Each itemset has an associated measure of statistical significance called *support.* For an itemset $X \subset \mathcal{I}$, *support(X)* $= s$, if the fraction of transactions in $\mathcal{D}$ containing X equals s. A rule has a measure of its strength called *confidence* defined as the ratio *support(X* $\cup$ *Y) / support(X).*

The problem of mining association rules is to generate all rules that have support and confidence greater than some user specified minimum support and minimum confidence thresholds, respectively. This problem can be decomposed into the following subproblems:

1. All itemsets that have support above the user specified minimum support are generated. These itemset are called the *large* itemsets. All others are said to be *small.*

2. For each large itemset, all the rules that have minimum confidence are generated as follows: for a large itemset X and any $Y \subset X$, if *support(X)/support(X* $-$ *Y)* $\geq$ *minimum_confidence*, then the rule $X - Y \longrightarrow Y$ is a valid rule.

To reduce the combinatorial search space, all algorithms exploit the following property, called *antimonotonicity* [11]: whenever the support of a set S of items violates the frequency constraint (i.e., the support falls below the specified threshold), then all supersets of S must also violate the frequency constraint. Conversely, any subset of a large itemset must also be large. This property is used by all existing algorithms for mining association rules (e.g., the Apriori algorithm [2]) as follows: initially support for all itemsets of length 1 (1-itemsets) are tested by scanning the database. The itemsets that are found to be small

are discarded. A set of 2-itemsets called *candidate itemsets* are generated by extending the large 1-itemsets generated in the previous pass by one (1-extensions) and their support is tested by scanning the database. Itemsets that are found to be large are again extended by one and their support is tested. In general, some kth iteration contains the following steps:

1. The set of candidate k-itemsets is generated by 1-extensions of the large $(k-1)$-itemsets generated in the previous iteration.
2. Supports for the candidate k–itemsets are generated by a pass over the database.
3. Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k-itemsets.

This process is repeated until no more large itemsets are found.

Other recent work [4, 5, 10] deals with finding rules based on other metrics besides support and confidence. In [4], the authors mine association rules that identify correlations and consider both the absence and presence of items as a basis for generating the rules. The measure of significance of associations that is used is the *chi-squared test* for correlation from classical statistics. In [5], the authors still use support as part of their measure of interest of an association. However, when rules are generated, instead of using confidence, the authors use a metric they call *conviction*, which is a measure of implication and not just co-occurrence. In [10], the authors also look at alternative measures of interest, namely the *gini index*, *entropy gain* and *chi-squared*. The problem examined in [10] is to find association rules that segment large categorical databases into two parts which are optimal according to some objective function. The functions used are information-theoretic measures which are used to indicate the extent of which the divided data distribution differs from the original data distribution. In [3], the notion of mining optimized rules is presented where the authors show that rules which satisfy a number of different interest metrics such as support. confidence, entropy, chi-squared and conviction reside along a support/confidence border. Hence, mining rules along this border will retrieve rules satsfying all the above metrics.

Bond also bears some relationship to the notion of mining association rules with multiple minimum supports [8]. In [8], the authors present an approach to the *rare item problem*. The dilemma that arises in the *rare item problem* is that searching for rules that involve infrequent (i.e., rare) items requires a low support but using a low support will typically generate many rules that are of no interest. Using a high support typically reduces the

number of rules mined but will eliminate the rules with rare items. The authors attack this problem by allowing users to specify different minimum supports for the various items in their mining algorithm. So, frequent items may have high support and infrequent items low support. They generate large itemsets with possible combinations of frequent and rare items based on their *sorted closure property*. As we will see, with our metric of *Bond*, we will also be able to find infrequent associations that may be interesting to the user by specifying one minimum threshold value (i.e., the minimum *Bond* value).

In this paper we concentrate on finding associations but with a different slant. That is, we take a different view of significance. Instead of *support* we use what we call *bond*. We claim that this is another measure of significance that has its place in mining associations that have high interest value.

Bond is a measure of the interestingness of an association. It indicates the degree to which items in an association are related to each other. It is similar to the notion of how terms in documents are related to each other (e.g., in information retrieval systems). It is similar to support but with respect to a subset of the data rather than the entire data set. This has similarities to the work in [14] except in their work they define data subsets based on the data satisfying certain time constraints. The idea is to find all itemsets that are frequent in a set of user-defined time intervals. In our case, the characteristics of the data define the subsets not the end-user.

For example, consider a medical application where we have n patients, a small number of those patients, ρ, exhibit any of the 3 symptons X, Y and Z and a number of those patients, ϵ, exhibit all 3 symptons X, Y and Z. It may be that ϵ/n is lower (even much lower) than the minimum support needed to produce an association between X, Y and Z. Hence that association would be deemed uninteresting. However, a physician may still be interested in that association if ϵ is close to ρ, that is ϵ/ρ is greater than or equal to some minimum value. The relationship of ϵ/ρ is what we call *bond*. For example, there may be 10000 patients where 5 of those patients exhibit a specific set of symptons, S. It may also be that the number of patients that exhibit any of those specific symptons S is 10. The support for an association containing the symptons in S would only be 0.0005. However, the bond would be 0.5. In the next section we formally define bond and its properties.

2 Bond as a Measure of Interestingness

In this section we present a formal definition of bond and prove a number of properties about it.

We previously defined the set of m items $\mathcal{I}$ as $\{i_1, i_2, \ldots, i_m\}$ and the set of variable length transactions over $\mathcal{I}$ as $\mathcal{D}$. Each transaction contains a set of items which are a subset of $\mathcal{I}$. We further denote the set of transactions that contain item i_j as $\mathcal{D}_{i_j}$.

Definition: The *bond* of a set of items, $\mathcal{L}$ is

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}$$

We should note that the bond of $\mathcal{L}$ where $|\mathcal{L}| = 1$ is 1. As an example, consider a database with the following transactions (also shown in Table 1) $T_1 = \{A, B\}$, $T_2 = \{A, B, C\}$, $T_3 = \{C, D\}$, $T_4 = \{C, D\}$ and $T_5 = \{E, F\}$ where $\mathcal{I}$ is $\{A, B, C, D, E, F\}$. The support and bond for all itemsets with a nonzero support are shown in Table 2.

To be able to efficiently determine the itemsets that have a bond value greater than the minimum bond, we would like to be able to prune the space of possible itemsets. This was done with respect to support for the Apriori algorithm [2] which used the property that if a set of items is not a frequent itemset, then any superset of that set is not a frequent itemset. We will also prove that this *antimonotonicity* property with respect to bond holds. This will allow us to discard any itemset that does not meet the minimum bond threshold.

Theorem 1 *The* antimonotonicity *property holds with respect to bond. That is, If*

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|} \geq minbond$$

Then $\forall \mathcal{L}' \subset \mathcal{L}$

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|} \geq minbond$$

Proof: *(By contradiction) Given that*

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|} \geq minbond$$

Assume that $\exists \mathcal{L}' \subset \mathcal{L}$ *such that*

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|} < minbond$$

We can define $\mathcal{L}''$ *as* $\mathcal{L} - \mathcal{L}'$ *or* $\mathcal{L}$ *as* $\mathcal{L}' \bigcup \mathcal{L}''$ *where* $\mathcal{L}' \bigcap \mathcal{L}'' = \emptyset$. *So,*

1. $\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j} = (\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}) \bigcap (\bigcap_{\forall i_j \in \mathcal{L}''} \mathcal{D}_{i_j})$.
2. $\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j} = (\bigcup_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}) \bigcup (\bigcup_{\forall i_j \in \mathcal{L}''} \mathcal{D}_{i_j})$.

From (1), we know that $|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}| \geq |\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|$. *They will be equal when* $\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j} \subset \bigcap_{\forall i_j \in \mathcal{L}''} \mathcal{D}_{i_j}$ *and greater otherwise. From (2), we know that* $|\bigcup_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}| \leq |\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|$. *Therefore,*

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|} \geq \frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}$$

Since

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|} < minbond$$

then

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|} < minbond$$

which contradicts our known fact. □

Lemma 1 *The bond for a set of items, L, will be greater than or equal to the support for L. This can be seen directly from the definition of bond and support, where the bond(L) is*

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}$$

and the support(L) is

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\mathcal{D}|}$$

since

$$|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}| \leq |\mathcal{D}|$$

Lemma 1 provides some information about the relationship between bond and support but we would like to make that relationship more precise as shown in Theorem 2.

Theorem 2 *The greatest lower bound for support($\mathcal{L}$) for a set of items L that have bond($\mathcal{L}$) is $1/\mathcal{D}$.*

Proof: *Suppose we have a set of items L such that the bond($\mathcal{L}$) $\geq$ minbond. Substituting the definition for bond we get*

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|} \geq minbond$$

By multiplying both sides of the inequality by

$$\frac{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\mathcal{D}|}$$

we obtain the following

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\mathcal{D}|} \geq minbond \cdot \frac{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\mathcal{D}|}$$

Now we have the support of L on the left-hand side. Thus

$$support(\mathcal{L}) \geq minbond \cdot \frac{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\mathcal{D}|}$$

If we want to determine the smallest possible support for the association of items L, we need to find the lower bound for

$$|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|$$

which is

$$\max_{\forall i_j \in \mathcal{L}} |\mathcal{D}_{i_j}|$$

Hence, we have the following

$$support(\mathcal{L}) \geq minbond \cdot \frac{\max_{\forall i_j \in \mathcal{L}} |\mathcal{D}_{i_j}|}{|\mathcal{D}|}$$

This, however is not the greatest lower bound for support. In particular, we want the greatest lower bound for support of any association (i.e., any set of items). We could first see if the lowest possible nonzero support could be achieved for a particular minbond. The lowest nonzero support for any L would be $1/\mathcal{D}$. The question is whether this can be achieved and it indeed can as follows:

It is possible that all $i_j \in \mathcal{L}$ occur in only one transaction. We have

$$\max_{\forall i_j \in \mathcal{L}} |\mathcal{D}_{i_j}| = 1$$

In this case we would have a $bond(\mathcal{L}) = 1$. Since we are interested in the precise support for L we can replace $\geq$ with $=$ and minbond with the value for $bond(\mathcal{L})$ which yields

$$support(\mathcal{L}) = \frac{1}{|\mathcal{D}|}$$

Hence, the greatest lower bound for support, possible for any minbond value is in fact the lowest possible nonzero support.

□

We formalize the relationship between bond and confidence by way of Theorem 3.

Theorem 3 *The lower bound for the confidence of any rule produced from a set of items $\mathcal{L}$ such that $\mathcal{L}$ has bond($\mathcal{L}$) is minbond.*

Proof: *Suppose we have a set of items L such that the bond($\mathcal{L}$) $\geq$ minbond. Substituting the definition for bond we get*

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|} \geq minbond$$

By multiplying both sides of the inequality by

$$\frac{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|}$$

we obtain the following

$$\frac{|\bigcap_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|} \geq minbond \cdot \frac{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|}$$

where $\mathcal{L}' \subset \mathcal{L}$.

Now we have the confidence of $\mathcal{L}' \longrightarrow \mathcal{L} - \mathcal{L}'$ on the left-hand side. Thus

$$confidence(\mathcal{L}' \longrightarrow \mathcal{L} - \mathcal{L}') \geq minbond \cdot \frac{|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|}{|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|}$$

If we want to determine the smallest possible confidence for any association rule $\mathcal{L}' \longrightarrow \mathcal{L} - \mathcal{L}'$, we need the lower bound for

$$|\bigcup_{\forall i_j \in \mathcal{L}} \mathcal{D}_{i_j}|$$

which is

$$\max_{\forall i_j \in \mathcal{L}} |\mathcal{D}_{i_j}|$$

and the upper bound for

$$|\bigcap_{\forall i_j \in \mathcal{L}'} \mathcal{D}_{i_j}|$$

which is

$$\min_{\forall i_j \in \mathcal{L}'} |\mathcal{D}_{i_j}|$$

Hence, we have the following

$$confidence(\mathcal{L}' \longrightarrow \mathcal{L} - \mathcal{L}') \geq minbond \cdot \frac{\max_{\forall i_j \in \mathcal{L}} \mid \mathcal{D}_{i_j} \mid}{\min_{\forall i_j \in \mathcal{L}'} \mid \mathcal{D}_{i_j} \mid}$$

Since $\mathcal{L}' \subset \mathcal{L}$ we know that

$$\min_{\forall i_j \in \mathcal{L}'} \mid \mathcal{D}_{i_j} \mid \leq \max_{\forall i_j \in \mathcal{L}} \mid \mathcal{D}_{i_j} \mid$$

Hence, when

$$\frac{\max_{\forall i_j \in \mathcal{L}} \mid \mathcal{D}_{i_j} \mid}{\min_{\forall i_j \in \mathcal{L}'} \mid \mathcal{D}_{i_j} \mid} = 1$$

we have a lower bound for confidence of minbond.

□

To illustrate the relationship of Theorem 2 and Theorem 3 to a set of transactions, we examine the transactions shown in Table 1. We will use a *minbond* of 0.6, i.e. 60%. From Theorem 2, we know that any association that satisfies this minimum bond value will have a support of 1/5 at least. We also know, from Theorem 3, that any rule produced by an association with minimum bond will have a confidence of at least 0.6. If we examine Table 2, we see that there are 3 associations (of size greater than 1), that satisfy the minimum bond requirement. They are displayed in Table 3 along with their associated rules. One point to make is that just because an association has the lower bound for support and confidence, it does not necessarily satisfy the minbond requirement. In Table 2, all itemsets satisfy the lower bound for support but only the itemsets in Table 3 satisfy the minimum bond requirement. If we were to lower the minimimum bond to 0.5, we would still have the results shown in Table 3. However, the itemset $\{A, C\}$ would not only satisfy the equivalent lower bound for support (i.e., 0.2) but also the rule $A \longrightarrow C$ would satisfy the lower bound for confidence (i.e., 0.5). However, the itemset $\{A, C\}$ would not satisfy the minimum bond requirement of 0.5. Hence, generating associations and rules that satisfy the lower bound for support and confidence would not produce only associations and rules that satisfy the mimimum bond requirement. The bottom line is that the output from an association finding algorithm would be a superset of the solution but the exact subset (for the bond metric) can not be determined directly without having to make an additional pass over the transaction data.

	Items					
Transaction	A	B	C	D	E	F
T_1	1	1	0	0	0	0
T_2	1	1	1	0	0	0
T_3	0	0	1	1	0	0
T_4	0	0	1	1	0	0
T_5	0	0	0	0	1	1

Table 1: Set of 5 Transactions (Items per transaction indicated by a 1)

Itemset	Support	Bond
A	2/5	1
B	2/5	1
C	3/5	1
D	2/5	1
E	1/5	1
F	1/5	1
AB	2/5	1
AC	1/5	1/4
BC	1/5	1/4
CD	2/5	2/3
EF	1/5	1
ABC	1/5	1/4

Table 2: Comparison of Support and Bond for Itemsets

Association	Bond	Support	Rule	Confidence
$\{A, B\}$	1	0.4		
			$A \longrightarrow B$	1
			$B \longrightarrow A$	1
$\{C, D\}$	0.66	0.4		
			$C \longrightarrow D$	0.66
			$D \longrightarrow C$	1
$\{E, F\}$	1	0.2		
			$E \longrightarrow F$	1
			$F \longrightarrow E$	1

Table 3: Associations with bond $\geq$ 0.6 and their rules

```
procedure gen_large_itemsets
L_i = the set of large itemsets of length i
l_j = an individual candidate large itemset contained in L_i

1)  L_1 = {large 1-itemsets along with their tidlists}
2)  for (k = 2; L_k ≠ ∅; k++) do begin
3)     forall itemsets l_1 ∈ L_{k-1} do begin
4)        forall itemsets l_2 ∈ L_{k-1} do begin
5)           if l_1[1] = l_2[1] ∧ l_1[2] = l_2[2] ∧ ... ∧ l_1[k-1] < l_2[k-1] then
6)              c = l_1[1] · l_1[2] ··· l_1[k-1] · l_2[k-1]
7)              if c cannot be pruned then
8)                 c.tidlist = l_1.tidlist ∩ l_2.tidlist
9)                 c.unlist = l_1[1].tidlist ∪ ... ∪ l_1[k-1].tidlist ∪ l_2[k-1].tidlist
9)                 if | c.tidlist | / | c.unlist | ≥ minbond then
10)                   L_k = L_k ∪ {c}
11)       end
12)    end
13) end
14) return ∪_k L_k
```

Figure 1: Procedure gen_large_itemsets

3 Bond Algorithm

The main task of the Bond algorithm, shown in Figure 1, is to generate the large itemsets that satisfy the minimum bond requirement.

Associated with each itemset is a list, called the *tidlist*. The tidlist consists of all transaction identifiers of the transactions containing the itemset. Also associated with an itemset is the *union_tidlist*, (i.e. the set of transactions that contain any of the individual items in that itemset). The cardinality of the *tidlist* divided by the cardinality of the *union_tidlist* is the bond for the associated itemset. The bond for an extension of the itemset is determined as follows: suppose t_1 and t_2 are the tidlists associated with itemsets l_1 and l_2, and c_3 is an itemset obtained by extending l_1 with l_2 (as explained below). The bond for c_3 is given by the number of transactions that contain c_3 (i.e., the intersection) divided by the number of unique transactions that contain any item in c_3 (i.e., the union). The main computational difference in computing support versus bond is the cost of computing the *union_tidlist*.

For example, let $\{T_1, T_3, T_4\}$ be the list of transactions associated with itemset {1,2} and $\{T_1, T_4, T_7\}$ be the list associated with {1,3}. Now, the transactions that contain the candi-

date itemset {1,2,3} are given by the intersection of the lists of transactions associated with itemsets {1,2} and {1,3}, i.e., $\{T_1, T_4\}$. Let the tidlist for itemset {1} be $\{T_1, T_3, T_4, T_5, T_7\}$, the tidlist for itemset {2} be $\{T_1, T_3, T_4, T_6\}$ and the tidlist for itemset {3} be $\{T_1, T_4, T_7\}$. The bond for itemset {1,2,3} is the cardinality of the intersection of tids for {1,2} and {1,3} divided by the cardinality of the union of tids for {1}, {2} and {3}. If this satisfies the minimum bond then {1,2,3} is a large itemset.

Initially a 1–itemset is created for every item in the database. The tidlists for these itemsets are generated by reading the database. For all 1-extensions (2-itemsets) of these itemsets, the tidlist is generated by intersecting the tidlists of both the itemsets in the extension. For the 2-itemsets, the union of the 1-itemsets is simply computed as the sum of the counts of the two 1-itemsets minus the count of the 2-itemset. The 2-itemsets that do not satisfy the minimum bond are discarded. The remaining itemsets are the large itemsets. These itemsets are extended by 1 and the process is repeated. The extensions of the itemsets are created as follows: let l_1 and l_2 be two k–itemsets, containing $\{i_j, i_k, \ldots, i_m\}$ and $\{i_p, i_q, \ldots, i_t\}$ respectively. A 1–extension of l_1 (a $(k+1)$–itemset) is generated if the following condition is satisfied: $i_j = i_p \wedge i_k = i_q \wedge \ldots \wedge i_m \leq i_t$. The $(k+1)$–itemset consists of $\{i_j, i_k, \ldots, i_m, i_t\}$. This technique is similar to the candidate generation step described in [2].

For fast computation of the intersection, the tidlists are maintained as arrays and the sort–merge join algorithm is used. Recall that the *TID*s are in ascending order in the database. Hence the tidlists are in the sort order initially and all resulting tidlists are automatically generated in the sort order. This operation is of linear complexity on the length of the tidlist.

In our implementation, the tidlists of itemsets of length greater than 1 are not materialized. For example, to compute the support for {A,B,C,D}, the tidlists for A, B, C and D are intersected. No tidlist is generated for the itemset {A,B,C,D}. The advantage of this approach is that we need storage for the tidlists of only the 1-itemsets and hence the memory requirement can be estimated quite accurately.

The procedure `gen_large_itemsets` generates all large itemsets (of all lengths). The procedure is the same as used in our previous work [15] The prune step is performed as follows:

prune(c: k–itemset)
forall $(k-1)$–subsets s of c **do**
    if $s \notin L_{k-1}$ **then**

return "c can be pruned"

The prune step eliminates extensions of $(k-1)$-itemsets which are not found to be large, from being considered for calculating the bond. For example, if L_3 is found to be {{1,2,3}, {1,2,4}, {1, 3,4}, {1,3,5}, {2,3,4}}, the candidate generation initially generates the itemsets {1,2,3,4} and {1,3,4,5}. However, itemset {1,3,4,5} is pruned since {1,4,5} is not in L_3. This technique is same as the one described in [2] except in our case, as each candidate itemset is generated, its bond is determined immediately.

4 Performance Results

In this section we describe the experimental results of using our technique for generating associations with a minimum bond. We performed two sets of experiments, one using synthetic data and the other using a subset of the 1990 United States census data.

4.1 Synthetic Data

The synthetic data is generated such that it simulates customer buying patterns in a retail market environment. We have used the same basic method as described in [17]. All of the synthetic data sets consisted of 100,000 transactions taken over 1000 items. The data labeled *T10.I4* had an average transaction size of 10 and a maximum transaction size of 40. The data labeled *T20.I4* had an average transaction size of 20 and a maximum transaction size of 50. The data labeled *T10.I4Y* consisted of 99,900 transacations generated by the synthetic data generator and 100 additional transactions. Those 100 transactions were made up of subsets of 7 items which only appear in those 100 transactions. All of the 100 transactions contain the same 3 items and a random number of the remaining 4 items.

A comparison of the bond algorithm running times for data sets *T10.I4* and *T20.I4* is shown in Figure 2. The amount of data processed (in bytes) for *T20.I4* was approximately twice the amount of data processed for *T10.I4*. This was simply due to the larger average transaction size. In Figure 2, we see that the running time for each given data set was fairly constant, regardless of the bond value. The reason is that the number of large itemsets generated was small for any of the values in the range of desired bond values. For the *T10.I4* data set, the number of large itemsets ranged from 0 to 16. For the *T20.I4* data set, the number of large itemsets ranged from 0 to 64. If the number of large itemsets were to increase dramatically, we expect the running time to do so as well. This can be seen in the association finding algorithms that use support as well.

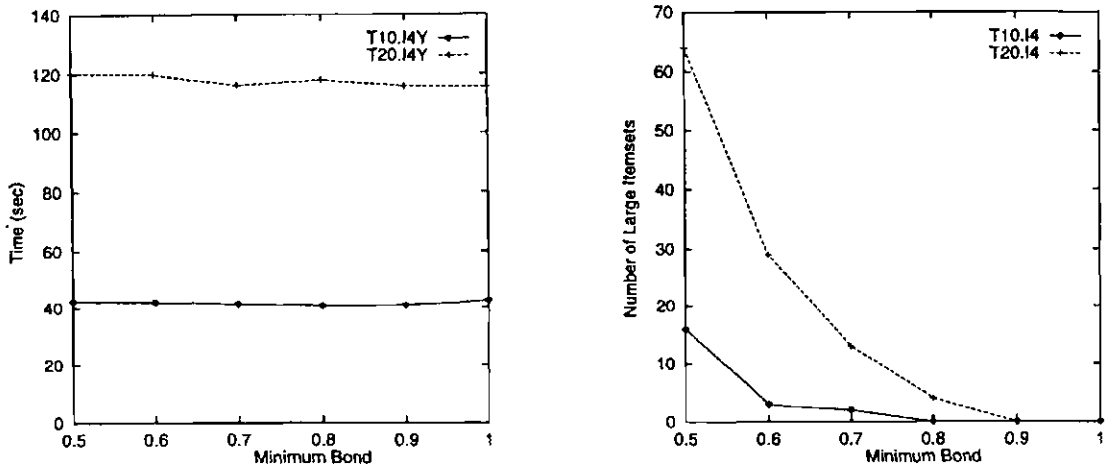

Figure 2: Performance of Bond Algorithm using synthetic data

Minimum	large itemset size							
	2		3		4		5	
Bond	sup	count	sup	count	sup	count	sup	count
1.0	0.1	3	0.1	1				
0.9	0.1	3	0.1	1				
0.8	0.1	3	0.1	1				
0.7	0.06	9	0.08	4	0.08	1		
0.6	0.04	11	0.08	4	0.08	1		
0.5	0.02	27	0.06	11	0.06	5	0.06	1

Table 4: Minimum support (in %) and count for large itemsets with minimum bond

The results of running the bond algorithm for data set *T10.I4Y* is shown in Figure 3. In these experiments, we intentionally placed sets of items in transactions so as to satisfy the bond requirement. The number of large itemsets varied from 4 for a bond of 1.0 to 44 for a bond of 0.5. Once again, since the number of large itemsets did not vary much, the running times remained fairly constant.

In Table 4, we show what the corresponding minimum support would be for the large itemsets that were determined based on bond. For a minimum bond value of 0.5, the algorithm determined 27 large itemsets of size 2, of which the minimum support was 0.02%.

4.2 Census Data

The data used in the next set of experiments was obtained from the U.S. Census Bureau through their online data extraction system available on the web at *www.census.gov/DES/*

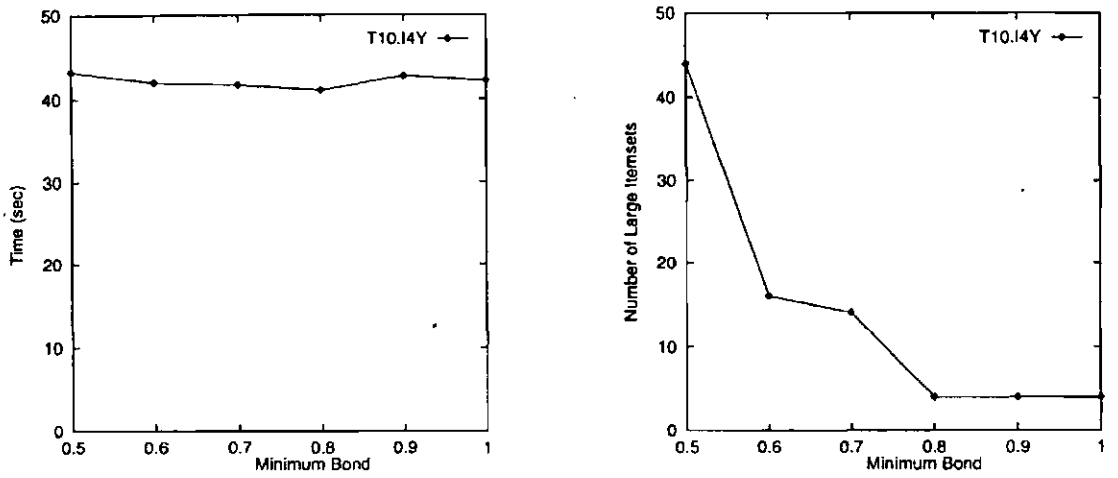

Figure 3: Performance of Bond Algorithm using synthetic data

www/welcome.html. The data is a subset of the 1990 Decennial Census Public Use Microdata 5% Samples. The data consisted of 53,847 records for people living in Florida of Hispanic origin. For these experiments we chose a subset of the available record fields, which included *age, citizenship, disability1, disability2, English, fertile, Hispanic origin, hours89, income1, language, martial, means, military, race, sex, year school and imigrated.* Since the fields were not all Boolean valued, we converted the numeric values into disjoint ranges and associated a unique field with each. The ranges were chosen based on the online summaries provided by the U.S. Census Bureau. The data was converted into 118 items but each record only contained a maximum of 20 items. The result of running the bond algorithm for this census data subset set is shown in Figure 4. The number of large itemsets varied from 4 for a bond of 1.0 to 102 for a bond of 0.5. In these experiments, the running time for the different bond values was not relatively constant (as with the synthetic data) since the number of large itemsets increased more with a lower bond value.

If we examine the associations produced for a minimum bond value of 1.0, and look at the largest association (i.e., size 3) produced, we see that it includes the following items: *work limitation status is not applicable, person is less than 16 years of age AND work prevention status is not applicable, person is less than 16 years of age AND military service is not applicable, person is less than 16 years of age.* These 3 items appeared in 11,427 records out of the 53,847 records.

If we examine some of the associations produced for a lower minimum bond value, we find somewhat more interesting associations. For example, with a minimum bond value

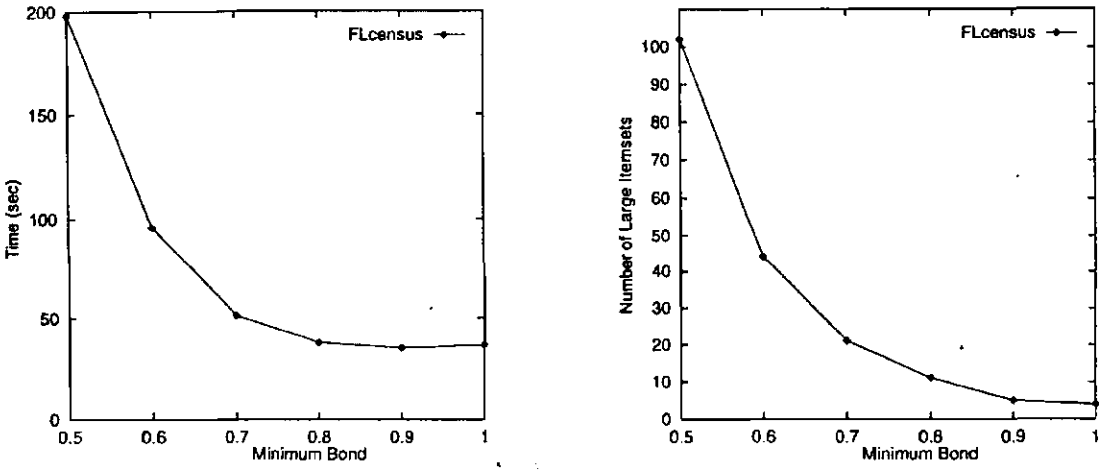

Figure 4: Performance of Bond Algorithm using U.S. census data

of 0.7, one association that was found was the following: *not limited from working AND not prevented from working AND speaks another language.* For a mimimum bond value of 0.5, some of the associations included *Hispanic origin is Puerto Rican AND born in Puerto Rico.* About half of the people of Puerto Rican origin were born in Puerto Rico. A corresponding association was not found for persons of other Hispanic origin such as Mexican or Cuban. Another sample association was *Hispanic origin is Cuban AND speaks another language.* Of the 32,934 persons of Cuban origin and the 45,000 people that speak another language, 29,709 persons speak another language and are of Cuban origin. One final sample association (of 5 items) for a mimimum bond of 0.5 was the following: *not limited from working AND not prevented from working AND speaks another language AND has no military service AND has immigrated to the U.S..*

5 Conclusion

In this paper we introduce an alternative measure of interestingness called bond. We prove that the important *antimonotonicity* property related to support also applies to bond. We also prove that if associations have a minimimum bond, then those associations will have a given lower bound on their mimimum support and the rules produced from those associations will have a given lower bound on their minimum confidence as well. We describe the algorithm that efficiently finds all associations with a minimum bond and present some experimental results, using both synthetic data and real-world census data. The performance

results indicate that the algorithm can find large itemsets with respect to the bond measure efficiently.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, DC, May 26-28 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, August 29-September 1 1994.

[3] R. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proceedings of the KDD Conference*, pages 145 – 154, August 1999.

[4] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD Conference*, pages 265 – 276, May 1997.

[5] S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD Conference*, pages 255 – 264, May 1997.

[6] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proceedings of the VLDB Conference*, pages 420 – 431, September 1995.

[7] M. Houtsma and A. Swami. Set-oriented mining of association rules. In *Proceedings of the International Conference on Data Engineering*, Taipei, Taiwan, March 1995.

[8] B. Liu, W. Hsu, and Y. Ma. Mining association rules with multiple minimum supports. In *Proceedings of the KDD Conference*, pages 337 – 341, August 1999.

[9] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. In *KDD-94: AAAI Workshop on Knowledge Discovery in Databases*, pages 181 – 192, Seattle, Washington, July 1994.

[10] Y. Morimoto, T. Fukuda, H. Matsuzawa, T. Tkuyama, and K. Yoda. Algorithms for mining associations rules for binary segmentation of huge categorical databases. In *Proceedings of the VLDB Conference*, pages 380 – 391, September 1998.

[11] R. Ng, L. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained associations rules. In *Proceedings of the ACM-SIGMOD Conference on Management of Data*, pages 13– 24, Seattle, Washington, June 1998.

[12] J. S. Park, M-S. Chen, and P. S. Yu. An effective hash based algorithm for mining association rules. In *Proceedings of the ACM-SIGMOD Conference on Management of Data*, pages 229 – 248, San Jose, California, May 1995.

[13] G. Piatetsky-Shapiro and W. J. Frawley, editors. *Knowledge Discovery in Databases.* MIT Press, 1991.

[14] S. Ramaswamy, S. Mahajan, and A. Silbershatz. On the discovery of interesting patterns in association rules. In *Proceedings of the VLDB Conference*, pages 368 – 379, September 1998.

[15] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules. In *Proceedings of the VLDB Conference*, pages 432 – 444, Zurich, Switzerland, September 1995.

[16] A. Savasere, E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of customer transactions. In *Proceedings of the IEEE Data Engineering Conference*, February 1998.

[17] R. Srikant and R. Agrawal. Mining generalized association rules. In *Proceedings of the VLDB Conference*, pages 407 – 419, September 1995.

Discovering Association Rules based on Image Content

Carlos Ordonez and Edward Omiecinski *
College of Computing
Georgia Institute of Technology
Atlanta, Georgia 30332-0280

December 7, 1998

Abstract

Our focus for data mining in this paper is concerned with knowledge discovery in image databases. In this first work we concentrate on the problem of finding associations. To that end, we present a data mining algorithm to find association rules in 2-dimensional color images. The algorithm has four major steps: feature extraction, object identification, auxiliary image creation and object mining. Our emphasis is on data mining of image content without the use of auxiliary domain knowledge. The purpose of our experiments is to explore the feasibility of this approach. A synthetic image set containing geometric shapes was generated to test our initial algorithm implementation. Our experimental results show that there is promise in image mining based on content. We compare these results against the rules obtained from manually identifying the shapes. We analyze the reasons for discrepancies. We also suggest directions for future work.

1 Introduction

Discovering knowledge from data stored in typical alphanumeric databases, such as relational databases, has been the focal point of most of the work in database mining. However, with advances in secondary and tertiary storage capacity, coupled with a relatively low storage cost, more and more non standard data (e.g., in the form of images) is being accumulated. This vast collection of image data can also be mined to discover new and valuable knowledge. The problem of image mining combines the areas of content-based image retrieval, image understanding, data mining and databases. This is a first attempt to combine association rules and images, but we know there has been significant research in image understanding in the Computer Vision community. An initial step towards tapping into the undiscovered wealth of knowledge from mining image-bases is the focus of this paper and more so, whether or not this is feasible.

There is a trend towards mining nonstandard and multimedia data [7]. Digitized images can be considered as a type of multimedia data. Current knowledge discovery technology is far from being able to extract all the knowledge contained in such diverse data types.

As related work to mining image content we can mention the following. There in an interesting prototype from Simon Fraser University called Multimedia Miner [16]. One of its modules is called MM-Associator. This module obtains association rules which are more restricted and simpler than the ones we obtain; these

*This work was supported in part by grant LM 06726-02 from the National Library of Medicine

rules relate information about the size, the color and the description of the image, but they do not involve specific objects identified automatically. Another important system used for discovering knowledge in a set of images is the Sky Image Cataloging and Analysis Tool (SKICAT) [6]. This program is used to study astronomical images. SKICAT uses trees and statistical optimization to classify objects obtained from an image segmentation process.

Image mining has two main themes. The first is mining large collections of images and the second is the combined data mining of large collections of image and associated alphanumeric data. As of now we have concentrated on mining only images; but our algorithm can be extended in a straightforward manner to handle images and associated alphanumeric data. An example of the first case might involve a collection of weather satellite imagery of various cities in the United States that has been recorded over an extended period of time. The data mining objective might be to find if there is some pattern that exists for an individual city (over time) or if there is some pattern that exists between different cities. An example of the second case might involve medical imagery and patient (alphanumeric data) records. To develop an accurate diagnosis or prognosis both image data (such as Xrays, SPECT, etc.) and patient data (such as weight, prior health conditions, family history, etc.) can be examined together to find interesting associations.

Our data mining system is built on top of a content-based image retrieval system (CBIR), the "Blobworld" system from the University of California at Berkeley (UCB). This CBIR system supports object-based queries and thus eliminates the need for the manual indexing of image content. This is a major advantage since manually indexing massive collections of images is impractical and prone to errors. Although CBIR systems are prone to retrieving non-related images.

2 Motivation

Applications such as military reconnaissance, weather forcasting, the management of earth's resources, criminal investigation and medical imaging all require (or will in the near future) the storage and processing of massive collections of images. For instance, NASA's EOS system generates 1 terabyte of image data per day. To take full advantage of these image-bases and the knowledge they contain, data mining techniques will have to be applied.

One of the typical data mining problems is to find association rules among data items in a database. In a retail environment such as a grocery store, an association rule might state that *customers who purchase spaghetti and Italian sausage also purchase red wine.* This can be determined by examining all the customer transactions (i.e., purchases). In this case, the data is explicit, there is a specific data item for each of the three grocery items and an individual customer transaction would include a subset of those items and in general a subset of all the items sold by the store. In the case of image-bases, assumming that all the images have been manually indexed (or their contents classified) may not be feasible. This presents one major deviation (problem) from the typical data mining approach for numerical data. If images can efficiently be labelled by a semantic descriptor, then the mining can be done on those high level concepts. However, with hundred's of thousands of images, this will become impossible. An alternative is to rely on automatic/semi-automatic analysis of the image content and to do the mining on the generated descriptors. For example, color, texture, shape and size can be determined automatically. Objects in an image can be determined by the similarity of those attributes. This is the approach we take in this first implementation.

3 Content-based Image Retrieval

Content-based image retrieval (CBIR) systems will be needed to effectively and efficiently use large image databases. With a CBIR system, users will be able to retrieve relevant images based on their contents. CBIR researchers have typically followed two distinct directions [10] based on

- modelling the contents of the image as a set of attributes which is produced manually and stored, for example in a relational database
- using an integrated feature-extraction/object-recognition system.

More recent research has recognized the need to develop a solution that captures the essence of both directions. However, there are still differences between the various current approaches. Mainly the differences can be categorized in terms of image features extracted, their level of abstraction and the degree of domain independence. Certainly tradeoffs must be made in building a CBIR system. For example, having automatic feature extraction is achieved at the expense of domain independence. Having a high degree of domain indepence is achieved by having a semiautomatic (or manual) feature extraction component.

With CBIR systems [10], querying is facilitated through generic query classes. Examples of some query classes include color, texture, shape, attributes, text and domain concepts. Color and texture queries allow users to formulate the description of the images to be retrieved in terms of like color and texture. Queries can also be posed with regard to the text associated with the images. For instance, in a medical setting, image retrieval is not only based on image content but also on the physician's diagnosis, treatment, etc. (i.e., additional textual data). We should also point out that CBIR differs from traditional database systems in that images are retrieved based on a degree of similarity and that records are usually retrieved from databases because of exactly matching specified attribute values.

Various content-based retrieval systems (QBIC, Chabot and Photobook) have focused on material-oriented queries and have used low-level image properties (e.g., color and texture) to implement these queries. On the other hand, a content-based retrieval system developed at the University of California at Berkeley [3, 4] focuses on object-oriented queries. That is, queries that search for images that contain particular objects. The approach to object recognition at Berkeley is structured around a sequence of increasingly specialized grouping activities that produces a "blobworld" representation of an image, which is a transformation from the raw pixel data to a small set of localized coherent regions in color and textual space. The "blobworld" representation is based on image segmentation using the Expectation-Maximization algorithm on combined color and texture features.

The salient feature of the Berkeley work is their approach to object recognition. Their approach [3] is based on the construction of a sequence of successively abstract descriptors through a hierarchy of grouping and learning processes. The image descriptors at a low level are color, texture, contrast, polarity, etc. and the grouping is based on spatiotemporal coherence of the local descriptors. The central notion in grouping is coherence and four major issues have been identified in [8] which are segmenting images into coherent regions based on integrated region and contour descriptors; fusing color, texture and shape information to describe primitives; using learning techniques for developing the relationship between object classes and color, texture and shape descriptors; and classifying objects based on primitive descriptors and relationships between primitives.

4 Data Mining

Database mining, an important part of knowledge discovery, is defined as the automated discovery of previously unknown, nontrivial, and potentially useful information from databases. The information is a statement that describes the relationship among a set of objects contained in the database with a certain confidence such that the statement is in some sense simpler than enumerating all the relationships between the individual instances of objects [9]. For example, in a database of employees and their salaries, each instance represents the relationship between an individual employee and his salary. A statement such as "salaries of engineers is higher than the salaries of secretaries," based on the instances of the database, conveys information that is implicit and more interesting than listing the salaries of all engineers and secretaries. Database mining is the process of generating high-level patterns that have acceptable certainty and are also interesting from a database of facts.

Knowledge discovery derives much of its success from reasoning techniques in artificial intelligence, expert systems, machine learning and statistics. Many paradigms such as inductive learning [14], Bayesian statistics [11], mathematical taxonomy [5], etc, have also been applied to knowledge discovery. In general, knowledge discovery is an amalgamation of concepts from diverse fields.

Efficiency is, in general, important for any computational problem. However, for database mining it also determines whether a particular technique can be applied or not. For example, the number of possible ways to cluster N objects into m clusters in unsupervised learning is exponential in N [12]. Hence, an algorithm which uses exhaustive search for clustering is impractical for real-world databases. In general any algorithm which grows faster than $O(n^2)$ is unlikely to be useful for large databases. Over the years, database systems, mainly relational, have made great strides in improving efficiency. The success of relational database systems in the business community can be attributed to these improvements. Many techniques such as, efficient access methods, buffer management, disk management, etc, are well understood. However, most of these techniques have been developed for on-line transaction processing (OLTP) applications. The access patterns for OLTP applications, which typically access a few hundred records, are considerably different from database mining applications, where entire tables may need to be scanned. One of the challenges in database mining is developing more efficient algorithms, better access structures, optimizing disk I/O, and so on.

5 Mining Image Content

There are two major issues that will affect the image data mining process. One is the notion of similarity matching and the other is the generality of the application area, that is, the breadth of usefulness of data mining from a practical point of view. For a specific application area, associated domain knowledge can be used to improve the data mining task. Since data mining relies on the underlying querying capability of the CBIR system, which is based on similarity matching, user interaction will be necessary to refine the data mining process.

The essential component in image mining is identifying similar objects in different images. With typical basket-market analysis, the data is usually constrained to a fixed set of items that are explicitly labelled. It is also quite efficient to see if a transaction contains a particular item, i.e., requires an examination of the item labels associated with a transaction. In some cases the data might be pre-processed into a fixed record format where a field exists for each item in the domain and a Boolean value is associated with it, indicating the presence or absence of that item in the transaction. This preprocessing can be done automatically. In a

general image mining setting, having a human label every possible object in a vast collection of images is a daunting task. However, we intend to capitalize on the recent work in CBIR, in particular, Blobworld [8].

We built our data mining system on top of a content-based image retrieval system. One premise behind supporting object-based queries in a CBIR system is to eliminate the need for manual indexing of image content. The CBIR system we use is from Berkeley [8]. We will refer to it as the "Blobworld" system. This system produces a "blobworld" representation of each image. A "blob" is just a 2-D ellipse which possesses a number of attributes. An image is made up of a collection of blobs, usually less than ten. Each blob represents a region of the image which is relatively homogeneous with respect to color and texture. A blob is described by its color, texture and spatial descriptors. The descriptors are represented by multidimensional vectors. Most of our limitations stem from the quality of the representation of image content by Blobworld.

At this point, we will consider in detail, the problem of finding associations. The problem of generating association rules was first introduced in [1] and an algorithm called *AIS* was proposed for mining all association rules. In [13], an algorithm called *SETM* was proposed to solve this problem using relational operations. In [2], two algorithms called *Apriori* and *AprioriTid* were proposed. These algorithms achieved significant improvements over the previous algorithms. The rule generation process was also extended to include multiple items in the consequent and an efficient algorithm for generating the rules was also presented. In [15], we presented an efficient algorithm for mining association rules that was fundamentally different from prior algorithms. Compared to previous algorithms, our algorithm not only reduced the I/O overhead significantly but also had lower CPU overhead for most cases.

Definitions

- *Association Rule.* An *association rule* is an implication of the form $X \Longrightarrow Y$, where $X, Y \subset \mathcal{I}$, and $X \cap Y = \emptyset$. $\mathcal{I}$ is the set of objects, also referred to as items. X is called the antecedent and Y is called the consequent of the rule. In general, a set of items, such as the antecedent or the consequent of a rule, is called an *itemset*.
- *Support.* Each itemset has an associated measure of statistical significance called *support*. For an itemset $X \subset \mathcal{I}$, *support*$(X) = s$, if the fraction of records in the database containing X equals s.
- *Confidence.* A rule has a measure of its strength called *confidence* defined as the ratio *support*$(X \cup Y)$ / *support*(X).

Algorithm to mine association rules

The problem of mining association rules is to generate all rules that have support and confidence greater than some user specified minimum support and minimum confidence thresholds, respectively. This problem can be decomposed into the following subproblems:

1. All itemsets that have support above the user specified minimum support are generated. These itemset are called the *large* itemsets. All others are said to be *small*.
2. For each large itemset, all the rules that have minimum confidence are generated as follows: for a large itemset X and any $Y \subset X$, if *support*(X)/*support*$(X - Y) \geq$ *minimum_confidence*, then the rule $X - Y \Longrightarrow Y$ is a valid rule.

6 Image Mining Algorithm steps

In this section, we present the algorithms needed to perform the mining of associations within the context of images. The four major image mining steps are as follows:

1. Feature extraction. Segment images into regions identifiable by region descriptors (blobs). Ideally one blob represents one object. This step is also called segmentation.
2. Object identification and record creation. Compare objects in one image to objects in every other image. Label each object with an id. We call this step the preprocessing algorithm.
3. Create auxiliary images. Generate images with identified objects to interpret the association rules obtained from the following step (html page creation).
4. Apply data mining algorithm to produce object association rules.

Here we explain the Image Mining processing in more detail. We keep I/O at a minimum. Images are kept on disk. For feature extraction each image is accessed once. These features are stored in two files, one is an image with all the blobs and the other with the blob descriptors. These blob descriptors are used to build an array with all the features from all the images. Once features are extracted from images we perform object identification using only their blob descriptors; this process is performed entirely in memory. Auxiliary images are kept on disk; these images show each identified object.

Images are not indexed because it is not necessary to search their contents once thay are segmented. Arrays of records are all that are needed to mine images once we have their features. Processing each image is performed independently of each other for feature extraction and this is done sequentially.

Identified objects, object associations and association rules are stored in sequential text files for interpreting results but not for processing. Our program can work with a large number of transactions. It is only limited by the amount of memory occupied by discovered associations.

Segmentation Step

It is not our intention to describe in detail the feature extraction process from the blobworld system. We will rather outline the main steps involved in identifying coherent image regions.

1. Estimate scale color selection σ.
2. Produce 6-dimensional feature vectors. These vectors contain summary information about color and texture only.
3. Produce several clusterings of feature vectors using the Expectation Maximization (EM) method. The 2 dominant colors are determined here. The number of groups in each clustering is called K.
4. Use the Minimum Description Length principle to decide which is the best K.
5. Segment the image into K regions using the spatial grouping of pixels. Each region is connected.
6. Apply a 3x3 max-vote filter to determine dominant colors.

INPUT: n segmented images, $\{I_1, I_2, \ldots, I_n\}$,
where I_i is a record containing: an image id and a blob descriptor vector bd
OUTPUT: Set of n records, $\{R_1, R_2, \ldots, R_n\}$ containing the object identifiers for the blobs

```
FOR i1 = 1 to n DO
  R_i1 = ∅
ENDFOR object_id = 0
FOR i1 = 1 TO n − 1 DO
  FOR j1 = 1 TO size(I_i.bd)
    first_time = true
    FOR i2 = i1 + 1 TO n
      IF I_i2.bd_j2 is not matched yet THEN
        IF similar(I_i1.bd_j1, I_i2.bd_j2, similarity_threshold, standard_deviation) THEN
          IF first_time THEN
            object_id = object_id + 1
            first_time = false
          ENDIF
          R_i1 = R_i1 ∪ { object_id }
          R_i2 = R_i2 ∪ { object_id }
          Mark I_i2.bd_j2 as matched
        ENDIF
      ENDIF
    ENDFOR
    Mark I_i1.bd_j1 as matched if there was one match at least
  ENDFOR
ENDFOR
Filter out unwanted matched objects
```

Figure 1: Preprocessing Algorithm

7. Generate blobs with summary information about each region when such region has an area greater than 2% of the image area.

Each blob has the most important information about each region. This information includes color, texture, shape, area and position, but only the the first three are considered relevant.

Preprocessing Algorithm

The basic algorithm for finding associations between images/blobs is similar to our association finding algorithm *Partition* [15], as long as we preprocess the image data. By preprocessing the image data, we will identify and label objects contained in the images using the image query processing algorithm [4]. The output of the preprocessing step will be a set of records, $R_1, R_2, \ldots R_k$, one for each image, containing the object identifiers for the objects contained in the image. This step is quite intensive since it is a similarity search between images, actually image descriptors. However, once this is accomplished, the actual data mining step will not require the expensive similarity searching. Our preprocessing algorithm is shown in Figure 1.

First of all we initialize the n records, which store the object id's for each of the n images. This

algorithm has four nested loops. These loops are required to compare every blob against the rest in the image database. In this manner we do not miss any potential match. Note that comparisons are always made between blobs in different images. Assuming the blob descriptor vector dimension is bounded, then this algorithm is $O(n^2)$. This is a reasonable asumption since the segmentation step cannot produce a high number of blobs. Nevertheless, if the number of objects in each image can also grow without bound this algorithm is $O(m^2n^2)$ for m the number of possible different objects and n the number of images. This can render the algorithm slow if n and m are similar; in general this algorithm is fairly fast if $m << n$, which is the usual case in practice.

The variable $first\_time$ is used to generate new object id's when one blob is matched for the first time. This is necessary because a single object in one image may be similar to many other objects in the following images and all these objects should have the same id.

When one blob turns out to be similar to another one we add the object id to their corresponding records. The first is the record that is being compared against the rest and the second one is the record for which a match was found. The second object in the comparison will be discarded avoiding a future unnecessary comparison. The similarity function to compare blobs is expensive to compute as we will see. So this has an impact on the overall performance of the algorithm.

Each segmented image is treated as a record and its transformation to a set of identified objects will also produce a record. This representation will also give us a direct way to incorporate alphanumeric information associated with the image into the image mining process. The algorithm can handle such information without modification, provided those additional attributes are treated as boolean values.

Similarity Function

The similarity function [4] between two blobs is essential for our image mining program. This function takes four parameters, the two blobs to be compared, a similarity threshold and a vector of standard deviations. The similarity function is mathematically defined as:

$$similarity = e^{-\frac{distance(blob_1, blob_2)}{2}},$$

where

$$distance(blob_1, blob_2) = [(blob_1 - blob_2)^T \Sigma^{-1} (blob_1 - blob_2)]^{1/2}.$$

In these formulas $blob_1$ and $blob_2$ are vectors containing summary features and Σ^{-1} represents the vector containing the standard deviations allowed for matching on each desired feature. The -1 power means we divide each distance by the corresponding entry of this vector; this is clarified in Figure 2. This similarity measure is 1 if there is a perfect match on all desired features and approaches zero as the match becomes worse. A low similarity can mean every object is similar to any other object.

It is important to note that the distance for color is computed in a special manner. Colors are stored as three coordinates for a point located in a color-cone, referred as the hue saturation value (hsv). The distance is computed as the minimum pairwise distance of the two dominant colors of the object. Each color is a point in 3-dimensional space affecting its third coordinate by a weight of 0.5 and leaving the first two unchanged. This is computed as a matrix product between the color vector and the weights. This distance constitutes

$INPUT: bd_1, bd_2,\ similarity\_threshold, standard\_deviation$
$OUTPUT: 1\ for\ a\ match,\ 0\ otherwise$

$d11 = (blob_1.cone_1 - blob_2.cone_1) * [\ 1\ 1\ 0.5]$
$d22 = (blob_1.cone_2 - blob_2.cone_2) * [\ 1\ 1\ 0.5]$
$d1 = |d11| + abs|d22|$
$d12 = (blob_1.cone_1 - blob_2.cone_2) * [\ 1\ 1\ 0.5]$
$d21 = (blob_1.cone_2 - blob_2.cone_1) * [\ 1\ 1\ 0.5]$
$d2 = abs|d12| + abs|d21|$
$dist_1 = min(d1, d2)$
$dist_{2:11} = blob_1.features - blob_2.features$
$score = exp(\ -0.5 * sqrt(\Sigma_{i=1}^{11}((dist_i/standard\_deviation_i)^2))\ )$

$return\ score >= similarity\_threshold$

Figure 2: Similarity function

the first entry of the difference vector. For all the remaining 10 features the distance is just computed as the difference between each blob entry.

The standard deviation vector permits adjusting parameters for the image mining process to use or discard features in an easy way. If we want to pay close attention to one specific feature we set the standard deviation to a value close to zero, but never zero. If we do not consider some feature to be relevant to the mining process we set the standard deviation to a high value. For some of the features the blobworld system requires standard deviations to be in specific ranges. More specifically standard deviations for color, anisotropy, and contrast require standard deviations at a maximum value of roughly 0.5. For area at most 0.1 is permitted. For all the remaining features any positive standard deviation is legal. If some feature is considered as completely irrelevant an infinity value is assigned to the corresponding entry.

A more detailed description of the similarity function between two blobs is given in Figure 2.

Auxiliary Image Creation

Here we talk about the auxiliary image creation step. It is important to mention that this step is necessary in order for the user to make sense out of the image mining results. We use a web browser as the tool to have an integrated view of images, image features (blobs), object ids, associations, and association rules.

The association rules are of the form:

$$\{id_1, id_2 \ldots id_k\} \Longrightarrow \{id_{k+1}, \ldots id_n\}, support = X\%, confidence = Y\%$$

For each image we show the original image with all the geometric shapes and then one blob image per matched blob. Ideally, each blob should correspond to one shape but this does not always happen as we will discuss. Each of of the blobs is labeled with the *object id* generated by the preprocessing algorithm. These are the id's that appear in the rule. Right now this step is somewhat slow because it involves generating one image per matched blob, but this process is done only once if the image mining program is run several times over the same image set. And also, this step is alleviated by the fact that unmatched blobs are not displayed and thus no image for them is generated.

After showing all the images there are two links to view the association text file and the association rule text file generated by the previous step. These files contain all the associations and rules as well as statistical information that help the user interpret and verify the correctness of the experimental results.

Example

At this point we show a simple example illustrating how our image mining algorithms work with $n = 10$. The original images and their corresponding blobs are shown on Figure 3 and Figure 4. Association rules corresponding to the identified objects are shown on Figure 5. We chose 10 representative images from the image set we created for our experiments. This set of synthetic images is explained in detail in the next section.

In Figures 3 and 4 we show the original image at the left with several geometric shapes and white background. These images are labeled with an image id. These images are the *only* input data for our program; no domain knowledge is used. Then we show a series of blob images, each containing one blob. These images are labeled with the id obtained in the preprocessing algorithm. Each blob has a close (most times equal) position to its corresponding geometric shape. There are some cases in which one blob corresponds to several geometric shapes.

For instance in image 004 the object 2 corresponds to the triangle. Object 1 is the background and is eliminated from consideration by filtering out the blobs corresponding to it. Note that in the image 033 the circle has two blobs corresponding to it (6 and 11).

Some of the blobs do not correspond to one shape, but to a set of shapes. This the case for blob 8 in image 103 and image 131, or the first blob 9 in image 119. It is important to note that these object identifiers are undesirable but the association rule algorithm eliminates them because they only appear in a couple of cases because their support is low. Also, it is important to note that object 9 is really the ellipse but there are no rules involving 9. Also object 2 is always the triangle except in image 108 in which it also corresponds to an ellipse. It is interesting that the ellipse gets such identifier because this only happened in this case; that is, there is no image in which the ellipse gets 2 and there is no triangle. This is again, a problem arising from the feature extraction step.

As parameters for object identification we set color standard deviation to 0.5, contrast standard deviation to 0.5 and anisotropy also to 0.5. The similarity threshold as needed by the similarity function was set to 0.6. We tuned these parameters after several experiments. These parameters maximized the number of associations and decreased the number of undesirable matches. The remaining parameters did not improve object identification for this set of synthetic images so their values were set to infinity.

The data mining algorithm was run with a 30% support and 70% confidence. The output is a set of association rules whose support and confidence are above these thresholds.

The 83 rules obtained by the program are shown on Figure 5. The rules we are going to explain here are marked with *** on the Figure 5. We preferred to show the entire output of our program in order to give the reader a truthful assesment of the results.

The first rule $\{4 \Rightarrow 2\}$ tells us that if there is an hexagon there is a triangle; the rule has confidence 1.0 and it is indeed correct. A similar rule is $\{10 \Rightarrow 2\}$ that says that if there is a square then there is a triangle. Note that these two rules have a correct confidence but their support is actually higher. This happens because the hexagon got two identifiers (4 and 5) and the square also got two identifiers (7 and 10). This also originates the problem of having repeated rules since $\{5 \Rightarrow 2\}$ is the same as $\{4 \Rightarrow 2\}$.

Now, looking at larger rules we see that the rule $\{4\ 7\ 11\} \Rightarrow \{2\}$ which says that an hexagon, a square and a circle imply a triangle is right. In fact, there is no image in this example in which these 3 shapes happen together and there is no triangle. A rule which has a lower confidence than 1 is $\{2\ 11\} \Rightarrow \{7\}$. The confidence for this rule is actually lower because image 029 has a triangle and a circle but it does not have a square, as implied by the rule; this happened because the circle in this image was identified as object 6 and was not also identified as object 11.

Some of the rules are redundant as is the case for the rule $\{5 \Rightarrow 4\}$. This rule says an hexagon implies an hexagon. An analog case is the rule for the square $\{10 \Rightarrow 7\}$. This problem is originated from the Blobworld system tht assigns two blobs to the same shape. However, in larger collections of images it may be the case that some rule can be discovered with one blob and not with the other one and therefore this might be helpful.

For this set of images none of the rules shown are false as can be verified. In some cases Their support or confidence are higher or a bit lower because the objects were incorrectly matched; but that difference is not significant. It is important to note that running the program with the lowest possible support (10%) does produce several incorrect rules since any rule valid for one image becomes valid for the entire set.

IMAGE MINING RESULTS

Number of mined images: 10

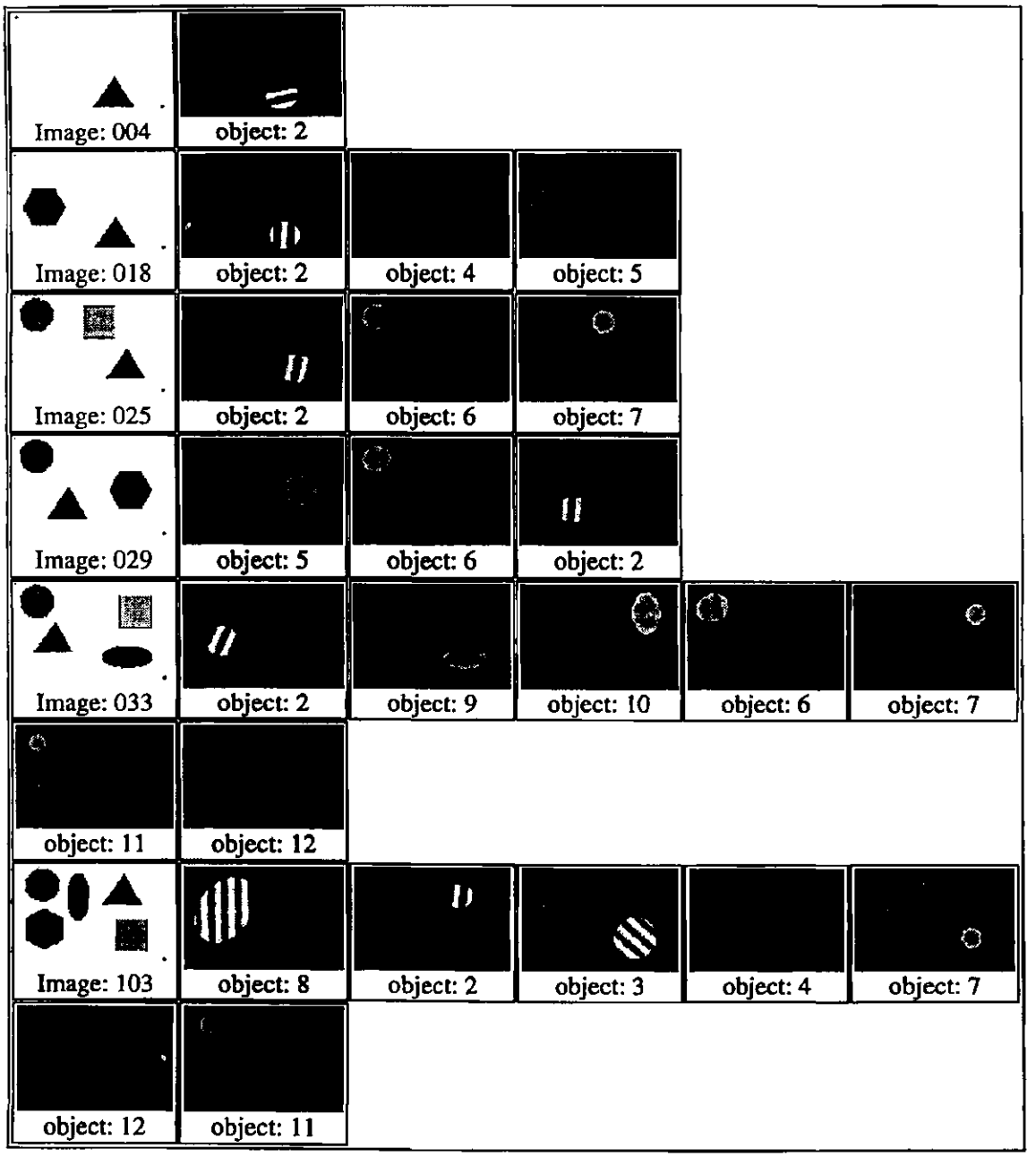

Figure 3: First part of images and blobs

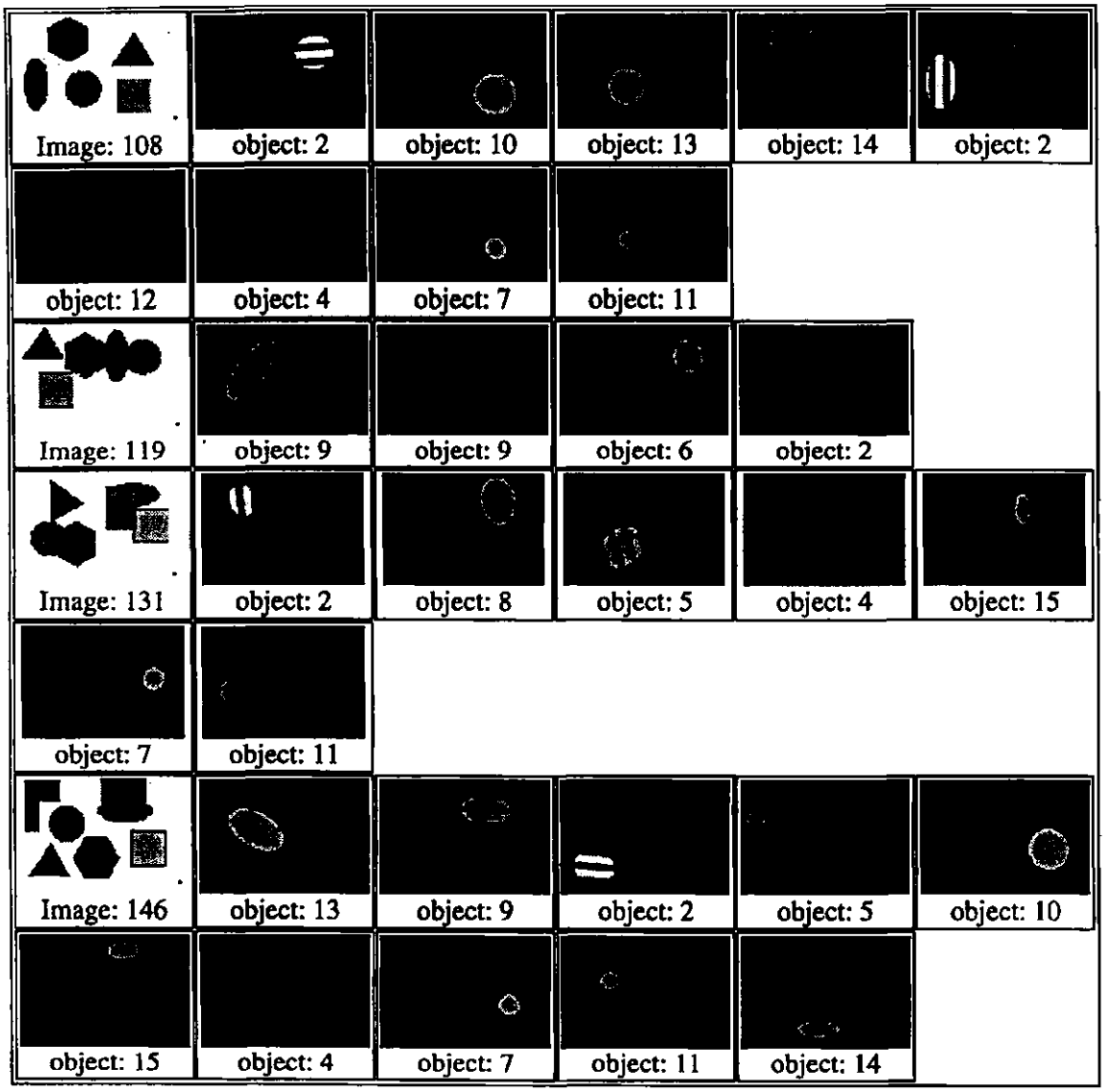

Figure 4: Second part of images and blobs

```
RULES GENERATED

Parameters:
Support:                30%
Confidence:             70%
Number of records:      10
Number of associations: 41
Support frequency:      3


{ 4 } => { 2 }   s= 0.50  c=1.00   ***     { 4 } => { 7 }   s= 0.40  c=0.80
{ 4 } => { 11 }   s= 0.40  c=0.80          { 4 } => { 2 7 }   s= 0.40  c=0.80
{ 4 } => { 2 11 }   s= 0.40  c=0.80        { 4 } => { 7 11 }   s= 0.40  c=0.80
{ 4 } => { 2 7 11 }   s= 0.40  c=0.80      { 5 } => { 2 }   s= 0.40  c=1.00  ***
{ 5 } => { 4 }   s= 0.30  c=0.75   ***     { 5 } => { 2 4 }   s= 0.30  c=0.75
{ 6 } => { 2 }   s= 0.40  c=1.00           { 7 } => { 2 }   s= 0.60  c=1.00
{ 7 } => { 11 }   s= 0.50  c=0.83          { 7 } => { 2 11 }   s= 0.50  c=0.83
{ 8 } => { 2 }   s= 0.30  c=1.00           { 9 } => { 2 }   s= 0.30  c=1.00
{ 10 } => { 2 }   s= 0.30  c=1.00  ***     { 10 } => { 7 }   s= 0.30  c=1.00  ***
{ 10 } => { 11 }   s= 0.30  c=1.00         { 10 } => { 2 7 }   s= 0.30  c=1.00
{ 10 } => { 2 11 }   s= 0.30  c=1.00       { 10 } => { 7 11 }   s= 0.30  c=1.00
{ 10 } => { 2 7 11 }   s= 0.30  c=1.00     { 11 } => { 2 }   s= 0.50  c=1.00
{ 11 } => { 4 }   s= 0.40  c=0.80          { 11 } => { 7 }   s= 0.50  c=1.00
{ 11 } => { 2 4 }   s= 0.40  c=0.80        { 11 } => { 2 7 }   s= 0.50  c=1.00
{ 11 } => { 4 7 }   s= 0.40  c=0.80        { 11 } => { 2 4 7 }   s= 0.40  c=0.80
{ 12 } => { 2 }   s= 0.30  c=1.00          { 12 } => { 7 }   s= 0.30  c=1.00
{ 12 } => { 11 }   s= 0.30  c=1.00         { 12 } => { 2 7 }   s= 0.30  c=1.00
{ 12 } => { 2 11 }   s= 0.30  c=1.00       { 12 } => { 7 11 }   s= 0.30  c=1.00
{ 12 } => { 2 7 11 }   s= 0.30  c=1.00     { 2 4 } => { 7 }   s= 0.40  c=0.80
{ 2 4 } => { 11 }   s= 0.40  c=0.80        { 2 4 } => { 7 11 }   s= 0.40  c=0.80
{ 2 5 } => { 4 }   s= 0.30  c=0.75         { 2 7 } => { 11 }   s= 0.50  c=0.83
{ 2 10 } => { 7 }   s= 0.30  c=1.00        { 2 10 } => { 11 }   s= 0.30  c=1.00
{ 2 10 } => { 7 11 }   s= 0.30  c=1.00     { 2 11 } => { 4 }   s= 0.40  c=0.80
{ 2 11 } => { 7 }   s= 0.50  c=1.00 ***    { 2 11 } => { 4 7 }   s= 0.40  c=0.80
{ 2 12 } => { 7 }   s= 0.30  c=1.00        { 2 12 } => { 11 }   s= 0.30  c=1.00
{ 2 12 } => { 7 11 }   s= 0.30  c=1.00     { 4 5 } => { 2 }   s= 0.30  c=1.00
{ 4 7 } => { 2 }   s= 0.40  c=1.00         { 4 7 } => { 11 }   s= 0.40  c=1.00
{ 4 7 } => { 2 11 }   s= 0.40  c=1.00      { 4 11 } => { 2 }   s= 0.40  c=1.00
{ 4 11 } => { 7 }   s= 0.40  c=1.00        { 4 11 } => { 2 7 }   s= 0.40  c=1.00
{ 7 10 } => { 2 }   s= 0.30  c=1.00        { 7 10 } => { 11 }   s= 0.30  c=1.00
{ 7 10 } => { 2 11 }   s= 0.30  c=1.00     { 7 11 } => { 2 }   s= 0.50  c=1.00
{ 7 11 } => { 4 }   s= 0.40  c=0.80        { 7 11 } => { 2 4 }   s= 0.40  c=0.80
{ 7 12 } => { 2 }   s= 0.30  c=1.00        { 7 12 } => { 11 }   s= 0.30  c=1.00
{ 7 12 } => { 2 11 }   s= 0.30  c=1.00     { 10 11 } => { 2 }   s= 0.30  c=1.00
{ 10 11 } => { 7 }   s= 0.30  c=1.00       { 10 11 } => { 2 7 }   s= 0.30  c=1.00
{ 11 12 } => { 2 }   s= 0.30  c=1.00       { 11 12 } => { 7 }   s= 0.30  c=1.00
{ 11 12 } => { 2 7 }   s= 0.30  c=1.00     { 2 4 7 } => { 11 }   s= 0.40  c=1.00
{ 2 4 11 } => { 7 }   s= 0.40  c=1.00      { 2 7 10 } => { 11 }   s= 0.30  c=1.00
{ 2 7 11 } => { 4 }   s= 0.40  c=0.80      { 2 7 12 } => { 11 }   s= 0.30  c=1.00
{ 2 10 11 } => { 7 }   s= 0.30  c=1.00     { 2 11 12 } => { 7 }   s= 0.30  c=1.00
{ 4 7 11 } => { 2 }   s= 0.40  c=1.00 ***  { 7 10 11 } => { 2 }   s= 0.30  c=1.00
{ 7 11 12 } => { 2 }   s= 0.30  c=1.00

Number of rules generated:   83
```


Figure 5: Association rules for identified objects

Range	Content
000-009	1 shape
010-019	2 shapes
020-029	3 shapes
030-049	4 shapes
100-109	5 shapes not overlapping without rectangle and L
110-119	5 shapes with some overlapping shapes
120-129	6 shapes not overlapping without L
130-139	6 shapes with some overlapping shapes
140-149	7 shapes, most complex images

Table 1: Image Content

7 Experimental Results

Synthetic Image Generation

To test our image mining algorithm we created synthetic images. We used synthetic images as a starting point in showing the feasibility of mining images. Also, with our constrained image set, we can more readily determine the weaknesses and strengths of our approach. These images are 192x128 color JPEG format because the Blobworld software from UCB needed this specific size. It is important to note that this format uses a lossy compression scheme and thus image quality is deteriorated. This was not a problem for our synthetic images, but it may be a problem for images with rich information content, such as photographs.

Our images contain a combination of plain geometric shapes. The seven shapes for our experiments were: triangle, circle, square, rectangle, hexagon, ellipse and an irregular shape similar to the letter L. Each of the shapes had a different uniform color and a black border. The background was always white. The texture for each shape was uniform with one exception, the irregular shape. For technical reasons two additional little objects were added to two opposing corners of each image to delimit its size. These little objects are ignored by the feature extraction step because it discards objects whose area does not represent more than 2% of the total image area.

Each geometric shape has the same size and color in each image where it appears. All shapes have a similar size with respect to each other. However, their position and orientation can differ between images. To make the mining process more interesting, in some cases we overlapped shapes or placed them very close to each other so they would seem to be part of the same object by the segmentation algorithm.

With the guidelines mentioned above we manually generated 100 basic images and we replicated some or all of these images to obtain larger image sets for our experiments. In Table 1 we show summary information for the images we created. We partitioned images into classes according to their content complexity. Image id's below 100 indicate easy to mine images and image id's greater than 100 mean difficult images, that is, images with complex content. In the easy images we have no more than 4 shapes plus the background. Shapes are in different positions but they are not close to each other and they are not overlapping either. For difficult images we have up to 7 shapes plus the background. In this case shapes overlap and also may be close to each other. This certainly makes the image mining process more difficult as we shall see. We want to stress our synthetic images are not as complex as images from the real world. This is a first attempt to mine association rules and thus we created simple images to carefully study the experimental results.

Category	Manual	Automatic
Number of associations	63	30
Number of rules	330	44
Maximum association size	6	4
Average association support	0.45	0.43
Average rule confidence	0.80	0.82

Table 2: Manual versus automatic image content mining

Hardware & Software

We ran our experiments on a Sun Multiprocessor (forge) computer with 4 processors (each running at 100 MHz) and 128 MB of RAM. The image mining program was written in Matlab and C.

The feature extraction process is done in Matlab by the software we obtained from UCB. Object identification and record creation were also done in Matlab. An html page is created in Matlab to interpret results. The asociation rules were obtained by a program written in C.

Quality of results

We should mention that there were no false association rules. It did not happen that an object was incorrectly identified and then a rule was generated with the incorrect identifier. In general when we found a match between two objects they were the same shape. All the incorrect matches are filtered out by the support parameter and then the association rules are generated for objects correctly ideintified. Also, some redundant matches happened because of the blobs that represented several shapes but these matches are filtered out by the rule support.

In the following table we present a summary of our experimental results with 100 hundred images. We compare the results obtained by manually identifying objects in each image and then generating association rules from such identifiers (Manual Column) against the results obtained by our current implementation (Automatic Column). Ideally, our image mining algorithm should produce the same results as the manual process. So, the table gives a standpoint to assess the quality of our experimental results. For these 100 images unwanted matches, either incorrect or involving many objects, happened in at most 4 images, and therefore their support was well below the minimum support frequency which was at 30.

These experiments were run using the same parameters for object identification as in our small example with 10 images. The parameters for object identification had the following values. We set color standard deviation to 0.5, contrast standard deviation to 0.5 and anisotropy also to 0.5. The similarity threshold as needed by the similarity function was set to 0.6. We tuned these parameters after several experiments. These parameters maximized the number of associations and decreased the errors in unwanted matches. The association rule program was set to look for rules with a 30% support and 70% confidence.

The background represents an object itself. Since association rules with the background were not interesting for our purposes it was eliminated from consideration by the object identification step. It is important to note that this is done after objects have been identified.

We tuned the object identification step to find similar objects changing values for several parameters

in the following manner. The most important features used from each object were color and contrast. We allowed some variance for color (0.5) and the maximum allowed variance for contrast (0.5). The anisotropy helped eliminate matches involving several geometric shapes. We ignored shape, because objects could be partially hidden and rotated. Position was considered unimportant because objects could be anywhere in each image. Anisotropy and polarity were ignored because almost all our shapes had uniform texture. Area was given no weight because objects could be overlapping, and thus their area diminished; this can be useful to make perfect matches when objects are apart from each other.

A few rules had high support. One problem that arose during our experiments was that the same shape could have two different blob descriptors, and these blob descriptors could not be matched with two other descriptors for the same shape in another image. This caused two problems. First, a rule could be repeated because it related the same shapes. Second, a rule did not have enough support and/or confidence and therefore was discarded. So, the rules found were correct and in many cases had an actual higher support and also higher confidence.

To our surprise in some cases there were no object matches because an object was very close to another one or was located in a corner of the image. When two or more objects were overlapping or very close they were identified as a single object. This changed the features stored in the blob. The problem was due to the ellipsoidal shape of the blobs and the fact that when a geometric shape was located in a corner thta changed its anysotropy and polarity descriptors. Given a blob for an object very close to one corner means determining an adequate radius for the blob (i.e., ellipse).

Regular shapes such as the triangle, square and hexagon were easily matched across images: This is a direct consequence of the circular blob representation produced when the image is segmented. In this case neither position nor rotation affect the mining process at all. It was surprising that in some cases there were no matches for the circle; in these cases it was in a corner or some other shape was very close or overlapping. Another important aspect about shape is that we do not use it as a parameter to mine images, but shape plays an important role during the segmentation step. So, shape does affect the image mining results quality.

The rectangle and the ellipse are the next shapes that are easily matched even though we did not use the shape feature. The most complicated shape was the L. In this case a number of factors affected matches. When this shape was overlapped with other shapes a few matches were found because a big blob was generated. Also, orientation changed dominant colors and contrast. When the L was close to another shape its colors were merged making it dissimilar to other L shaped objects. This suggests that irregular shapes in general make image mining difficult.

We worked with color images but it is also possible to use black and white images. Color and texture were important in mining the geometric shapes we created. However, we ignored shape as mentioned above. Shape may be more important for black and white images but more accurate shape descriptors, than those provided by the blobs.

Running time

Feature extraction is slow and there are several reasons for this. If image size increases performance should degrade considerably since feature extraction is quadratic in image size. Nevertheless, this step is done only once and does not have to be repeated to run the image mining algorithm several times. Object identification is fast. This is because the algorithm only compares unmatched objects and the number of objects per

image is bounded. For our experimental results time for this step scales up well. Auxiliary image creation is relatively slow but its time grows linearly since it is done on a per image basis. The time it takes to find rules is the lowest among all steps. If the image mining program is run several times over the same image set only the times for the *second* and the *fourth* step should be considered since image features already exist and auxiliary images have already been created.

Application

Image mining can have an application with real images. The current implementation can be used with a set of images having the following characteristics:

- Homogeneous. The images should have the same type of image content. For instance, the program can give useless results if some images are landscapes, other images contain only people and the remaining images have only cars.
- Simple image content. If the images are complex they will produce blobs difficult to match. Also, the association rules obtained will be harder to interpret. A high number of colors, blurred boundaries between objects, large number of objects, significant difference in object size make the image mining process more prone to errors.
- A few objects per image. If the number of objects per image is greater than 10 then our current implementation would not give accurate results since Blobworld in most cases generates at most 12 blobs per image.
- New information. The image itself should should give information not already known. If all the information about the image is contained in associated alphanumeric data, then that data could be mined directly.

8 Future Work

Results obtained so far look promising but we need to improve several aspects in our research effort. We are currently working on the following tasks.

We also need to analyze images with repeated geometric shapes. If we want to obtain simple association rules this can make our program more general. This can be done without further modification to what is working. However, if we want to mine for more specific rules then we would need to modify our algorithm. For instance, we could try to get rules like the following: if there are two rectangles and one square then we are likely to find three triangles. The issues are the combinatorial growth of all the possibilities to mine and also a more complex type of condition. We will also study more deeply the problem of mining images with more complex shapes such as the irregular one similar to the letter L.

We need a systematic approach to determine an optimal similarity threshold or at least a close one. A very high threshold means only perfect matches are accepted. On the other hand, a very low similarity threshold may mean any object is similar to any other object. Finding the right similarity threshold for each image type looks like an interesting problem. Right now it is provided by the user but it can be changed to be tuned by the algorithm itself. Also, there are many ways to tune the eleven parameters to match blobs and the optimal tuning may be specific to image type.

There also exists the possibility of using other segmentation algorithms that could perform faster or better feature extraction. It is important to note that these algorithms should give a means to compare segmented regions and provide suitable parameters to perform object matching in order to be useful for image mining. From our experimental results it is clear that this step is a bottleneck for the overall performance of image mining.

We can change the object identification algorithms to generate overlapping object associations using more features. Our algorithm currently generates partititons of objects, that is, if one object is considered similar To another one, the latter one will not be compared again. By generating overlapping associations we can find even more rules. For instance a red rectangular object may be considered similar to another rectangular object and at the same time be similar to another red object. Mining by position is also possible; for instance two objects in a certain position may imply another object to be in some other position. Since the software we are using for feature extraction produces eleven parameters to describe blobs we have 2^{11} possibilites to match objects.

9 Conclusions

We presented a new algorithm to perform data mining on images and an initial experimental and performance study. The positive points about our algorithm to find association rules in images and its implementation include the following. It does not use domain knowledge, it is reasonably fast, it does not produce meaningless or false rules, it is automated for the most part. The negative points include: some valid rules are discarded because of low support, there are repeated rules because of different object id's. unwanted matches because of blobs representing several objects, slow feature extraction step, a careful tuning of several parameters is needed.

We studied this problem in the context of data mining for databases. Our image mining algorithm has 4 major steps: feature extraction, object identification, auxiliary image creation and identified object mining. The slowest part of image mining is the feature extraction step, which is really a part of the process of storing images in a CBIR system; and is done only once. The next slowest operation is creating the auxiliary blob images which is also done once. Object identification and association rule finding are fairly fast and scale up well with image set size. We also presented several improvements to our initial approach of image mining.

Our experimental results are promising and show some potential for future study. Rules referring to specific objects are obtained regardless of object position, object orientation, and even object shape when one object is partially hidden. Image mining is feasible to obtain simple rules involving images with a few simple objects, but it needs human intervention and domain knowledge to get better results.

Images contain a great deal of information, and thus the amount of knowledge that we can extract from them is enormous. This work is an attempt to combine association rules with automatically identified objects obtained from a matching process on segmented images. Although our experimental results are far from perfect we show that it is better to discover some reliable knowledge automatically than not discovering any new knowledge at all.

Acknowledgments

We thank Chad Carson from the University of California at Berkeley for helping us setup the Blobworld system. We also thank Sham Navathe for his comments to improve the presentation of this paper.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, DC, May 26-28 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, August 29-September 1 1994.

[3] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Recognition of images in large databases using a learning framework. Technical Report TR 97-939, U.C. Berkeley, CS Division, 1997.

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Region-based image querying. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1997.

[5] G. Dunn and B. S. Everitt. *An Introduction to Mathematical Taxonomy.* Cambridge University Press, New York, 1982.

[6] U. Fayyad, , D. Haussler, and P. Storoltz. Mining scientific data. *Communications of the ACM*, 39(11):51–57, November 1996.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, November 1996.

[8] D. Forsyth, J. Malik, M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images. Technical report, U.C. Berkeley, CS Division, 1997.

[9] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. *Knowledge Discovery in Databases*, chapter Knowledge Discovery in Databases: An Overview, pages 1 – 27. MIT Press, 1991.

[10] V. Gudivada and V. Raghavan. Content-based image retrieval systems. *IEEE Computer*, 28(9):18–22, September 1995.

[11] R. Hanson, J. Stutz, and P. Cheeseman. Bayesian classification theory. Technical Report FIA-90-12-7-01, Artificial Intelligence Research Branch, NASA Ames Research Center, Moffet Field, CA 94035, 1990.

[12] M. Holsheimer and A. Siebes. Data mining: The search for knowledge in databases. Technical Report CS-R9406, CWI, Amsterdam, The Netherlands, 1993.

[13] M. Houtsma and A. Swami. Set-oriented mining of association rules. Technical Report RJ 9567, IBM, October 1993.

[14] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[15] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules. In *Proceedings of the VLDB Conference*, pages 432 – 444, Zurich, Switzerland, September 1995.

[16] O. R. Zaiane, J. Han, Z. N. Li, J. Y. Chiang, and S. Chee. Multimedia-miner: A system prototype for multimedia data mining. In *Proc. 1998 ACM-SIGMOD Conf. on Management of Data*, June 1998.

#3

PROGRESS REPORT SUMMARY	GRANT NUMBER	
	LM06726-04	
PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR	PERIOD COVERED BY THIS REPORT	
Norberto Ezquerra, Ph.D.	FROM	THROUGH
APPLICANT ORGANIZATION		
Georgia Institute of Technology	02/01/01	01/31/02

TITLE OF PROJECT *(Repeat title shown in Item 1 on first page)*
Knowledge Discovery in Distributed Cardiac Imagebases

a. Human Subjects *(Complete Item 7 on the Face Page)*
Use of Human Subjects ☐ Change ☒ No Change Since Previous Submission

b. Vertebrate Animals *(Complete Item 8 on the Face Page)*
Use of Vertebrate Animals ☐ Change ☒ No Change Since Previous Submission

(SEE INSTRUCTIONS)

Has there been a change in the other support of key personnel? No.
Will there be, in the next budget period, significant rebudgeting of funds? No.
Will there be a change in the level of effort for key personnel? No.
Is an unobligated balance expected? No.

Research Accomplishments

(A) SPECIFIC AIMS

The overall objective of the research program remains to discover potentially new knowledge regarding the assessment of coronary artery disease (CAD) by mining image and non-image databases both locally and from remote sites. The three specific aims have also remained the same. A number of publications have resulted from this work.

(B AND C) STUDIES, RESULTS, AND SIGNIFICANCE

Aim #1: Knowledge Discovery. There were several thrusts related to this aim:

(i) Knowledge-guided mining. Our association rule (AR) mining algorithms were placed in the hands of the clinical users who collaborate on the project. Scripts were created that capture specific mining goals and which exploit the users ŕ domain knowledge. These scripts were designed in the form of IF-THEN rule templates such that the antecedent and consequent contained fields or variables of particular medical interest. We used our existing knowledge-based system for interepreting stress perfusion imagery, PERFEX, as a guide to define meaningful rule templates. This knowledge-guided approach to mining is useful for several reasons: (a) it represents a novel mining approach, (b) this method results in a significant reduction in the number of association rules that are found, (c) the rules generated are based on statistically rigorous measures, and (d) the approach represents an innovative and efficient way to acquire potentially new knowledge associating a large number of variables. See [Ord00a; Ord00b].

(ii) Alternative interest measures for mining association rules. Interest measures are a statiscally-based way to determine the usefulness of the mining results. Three alternative measures were investigated: any-confidence, all-confidence, and bond. Several theoretical properties were discovered for these measures, and their potential impact on mining was examined. These efforts represent both basic computer science and practical contributions and were reported in a peer-reviewed journal article [Omi00].

(iii) Contrasts and Differences Between Datasets. We examined ways to find differences between different data sets. In general, it is important to detect and characterize possible differences between any number of data sets. This becomes useful to find differences between data sets obtained from different medical sites. There has been only limited work in this area. We have investigated a novel approach using AR mining which requires only 2 passes over each data set regardless of the number of items. Clustering algorithms were also explored [Ord00c; Ord00d].

Aim #2: Knowledge Base (KB) Enrichment

(i) KB confirmation and validation: The mining results were used to confirm pre-existing knowledge: i.e., knowledge rules that already appeared in the PERFEX KB. A dataset consisting of 655 patient cases was used for this study,

GENDER AND MINORITY INCLUSION
Provide the number of subjects enrolled in the study *to date* (cumulatively since the most recent competitive award) according to the following categories. (See Page 9 for definitions.) If there is more than one study, provide a separate table for each study. In addition, report on the subpopulations which are included in the study.

Study Title

	American Indian or Alaskan Native	Asian or Pacific Islander	Black, not of Hispanic Origin	Hispanic	White, not of Hispanic Origin	Other or Unknown	TOTAL
Female							
Male							
Unknown							
TOTAL							

which used coronary angiography as a gold standard in assessing arterial disease. Reported in [Gar00; San00a].
(ii) KB performance: A study was performed to compare the sensitivity and specificity (S&S) of PERFEX when using heuristically defined rule Certainty Factor (CF) values, with the S&S of PERFEX when these CF values were replaced with the values of support (statistical significance) derived from AR mining. The underlying assumption was that PERFEX is sufficiently robust when using the CF Model values. For this study, 655 patient cases were used and the comparison was performed only for left anterior descending disease. The S&S of both approaches were shown to be comparable to each other within the significance of the sample size. See [Coo00a; Coo00b; San00b].
(iii) Assessment of potentially new knowledge. We incorporated an important step in data mining to select clinically useful associations: the introduction of a truth table for each variable used in mining. Each table consists of all patient cases which are true positives, true negatives, false positives, and false negatives. Significantly, this effort provided a means with which to determine evidence against disease.

Aim #3: Distributed Knowledge Discovery

(i) Creation of Databases from remote sites. Collecting patient data from four remote collaborating sites has been emphasized during the current year and will continue during the upcoming year of research. Nearly 80% of the target 400 datasets were collected from the following sites: (a) St. Luke's (NYC), (b) Hospital Vall de'Hebron (Barcelona, Spain), (c) Cardiovascular Consultants (Kansas City, MO), and (d) Miami Baptist (FL). These datasets are comprised of all the relevant SPECT information and other non-image data.
(ii) Remote-site data preparation. One of the most important aspects of mining heterogeneous databases is the creation of mechanisms for creating formats such that uniformly support mining operations and yield meaningful interpretations of the results. There are numerous format differences between the clinical sites, and the possibility also exists of missing, misleading, incorrect, or mis-matching information in the data files. Hence, it became necessary to arrive at commonly accepted field definitions, identify a format optimal for mining, develop programs to convert formats into a common format, detect discrepancies, and perform a number of other low-level data operations.

(D) PLANS FOR NEXT YEAR OF SUPPORT

During the next project period, we will complete remote-site data preparation. We expect to gain valuable insights regarding knowledge discovery across different DBs, and will pursue the contrast-sets studies to examine differences. The thrust will be placed on comparing the efficiency of the different mining algorithms, mining the data from the different medical sites, continuing the knowledge-guided mining approach, and evaluating discovered knowledge mentioned above. Efforts will be made to disseminate the mining techniques and the conclusions reached thus far.

(E) PUBLICATIONS

1. [Coo00a] C.D. Cooke, C. Ordonez, E.V. Garcia, E. Omiecinski, E. Krawczynska, R. Folks, C. Santana, L deBraal, N. Ezquerra: Data Mining of large myocardial perfusion SPECT databases to improve diagnostics decision-making. J. Nucl. Med. 1999, Vol. 40., No. 5, p.292P (abstract) (2000).
2. [Coo00b] C. D. Cooke, C. Santana, T. Morris, L. de Braal, C. Ordonez, E. Omiecinski, N. Ezquerra, E. Garcia: Validating Expert System Rule Confidences Using Data Mining of Myocardial Perfusion SPECT Databases. Proceedings of Computers in Cardiology (in press). (2000)
3. [Gar00] E. Garcia, C. Cooke, R. Folks, C. Santana, E. Krawczynska, L. de Braal, N. Ezquerra: Diagnostic Performance of an Expert System for the Interpretation of Myocardial Perfusion SPECT Studies. Accepted for publication in the Journal of Nuclear Medicine.
4. [Omi00] E. Omiecinski: Alternative Interest Measures for Mining Association Databases, under second review by IEEE Tran. Knowledge and Data Engineering.
5. [Ord99] C. Ordonez and E. Omiecinski: Discovering Association Rules Based on Image Content. IEEE forum on Advances in Digital Libraries, ADL 1999, pp. 38-49
6. [Ord00a] C. Ordonez, C. Santana, L. de Braal: Discovering Interesting Association Rules in Medical Data, ACM SIGMOD Workshop on Data Mining and Knowedledge Discovery, pp. 78-85 (2000).
7. [Ord00b] C. Ordonez, N. Ezquerra, L. de Braal, C. Santana, E. Omiecinski, J. Taboada, E. Garcia, D. Cooke, E. Krawczynska: Mining Constrained Association Rules in Medical Data; to be submitted to IEEE Tran. Knowledge and Data Engineering.
8. [Ord00c] C. Ordonez, P. Cereghini: SQLEM: Fast Clustering in SQL Using the EM Algorithm, ACM SIGMOD Conference, pp. 559-570 (2000).
9. [Ord00d] C. Ordonez, E. Omiecinski: Building a Statistical Model for Association Rules, to be submitted to ACM SIGMOD 2001 Conference.
10. [Ord00e] C. Ordonez: Mining Complex Databases Using the EM Algorithm, Ph.D. Dissertation, Georgia Institute of Technology, Fall Semester 2000.
11. [San00a] C. Santana, E. Garcia, J. Vansant, E. Krawczynska, R. Folks, C. D. Cooke, T. Faber: Three-Dimensional Color Modulated Display of Myocaridal PSECT Perfusion Distributions Accurately Assess Coronary Artery Disease. Accepted for publication in the Journal of Nuclear Medicine.
12. [San00b] C. Santana, C. D. Cooke, M. Soler-Peter, T. Faber, E. Krawczynska, R. Folks, E. Garcia: Expert System (PERFEX) Interpretation of Rubidium-82 Myocardial Perfusion PET: Validation Using 53 Patients. Journal Nuclear Cardiology Vol. 7, No. 4, pS12 (abstract) (2000).